Information Retrieval und Informetrie: Zur Anwendung informetrischer Methoden in digitalen Bibliotheken

Philipp Schaer *

Abstract: »*Information Retrieval and Informetrics: The Application of Informetric Methods in Digital Libraries*«. The search for scientific literature in scientific information systems is a discipline at the intersection between information retrieval and digital libraries. Recent user studies show two typical weaknesses of the classical IR model: ranking of retrieved and maybe relevant documents and the language problem during the query formulation phase. At the same time traditional retrieval systems that rely primarily on textual document and query features are stagnating for years, as it could be observed in IR evaluation campaigns such as TREC or CLEF. Therefore alternative approaches to surpass these two problem fields are needed. Recent developments in the area of applied informetrics show very promising effects by using long-known informetric and bibliometric methods like the analysis of power-law distributions described by Lotka's, Zipf's or Bradford's laws. This contribution will concentrate on the description of the different approaches in digital libraries, information retrieval, and informetrics to give a broad overview on current methods in applied informetrics.This article contains:

1. Introduction
2. Digital Libraries
3. User Estimation of Relevance and Computer-Generated Ranking
4. Evaluation of Information Retrieval Systems
5. Informetrics
6. Discussion

Keywords: Digital libraries, informetrics, Power Law, Bradford's Law, Lotka's Law, Zipf's Law, information retrieval.

1. Einleitung

Digitale Bibliotheken, wissenschaftliche Literaturdatenbanken und Web-Informationssysteme sind unverzichtbare Rechercheplattformen für wissen-

* Philipp Schaer, GESIS – Leibniz-Institut für Sozialwissenschaften, Unter Sachsenhausen 6-8, 50667 Köln, Germany; philipp.schaer@gesis.org.

schaftliches Arbeiten geworden. Trotz vieler Systemneugründungen (Google Scholar oder Microsoft Academic Search) und bereits etablierter Systeme (arXiv und Web of Science) sind die grundlegenden Probleme bei der Suche meist die gleichen geblieben: Nutzer beklagen entweder einen „Information Overload“ oder das Phänomen „nichts zu finden“. Dieses grundsätzliche Problem ist auch als „feast or famine“ bekannt und begleitet die Fachdisziplin des Information Retrieval seit Anbeginn. Verschärft wird das Problem durch eine ausgeprägte „Now-or-Never-Mentalität“, die die Verfügbarkeit von Informationen in das Zentrum des Informationsbedürfnisses stellt: Was nicht in Sekundenbruchteilen gefunden werden kann, wird (fälschlicherweise) als nicht relevant eingestuft.

Nutzerstudien im Bereich der digitalen Bibliotheken und der wissenschaftlichen Informationsversorgung zeigen zwei besonders problematische Schritte im Retrievalprozess: Das Ranking potentieller Treffer und die Formulierung einer passenden Anfrage, um überhaupt zu diesem Treffer zu gelangen. Wie von Schaer (2013a) beschrieben werden diese grundsätzlichen Retrievalprobleme auch von modernen Retrievalsystemen nicht ausreichend gelöst, was sich auch an den stagnierenden Systemleistungen der großen IR-Evaluationskampagnen TREC und CLEF zeigt.

Ein Ausweg aus diesem grundsätzlichen aber trotzdem aktuellen Problem könnte die Erweiterung der Retrievalidee um nicht-sprachliche Elemente sein, wie sie z.B. informetrische Analysen bieten (Mutschke u.a. 2011). Während traditionell im Information Retrieval ein Abgleich zwischen Anfragetermen und Dokumenttermen im Zentrum des Retrievalprozesses steht, könnte die Analyse von Strukturen und Dynamiken zusätzliche Anhaltspunkte für einen erfolgreichen Retrievalprozess liefern. So wurde beispielsweise erst durch den Gedanken, die Relevanz von Webseiten nicht (primär) nach sprachlichen Eigenschaften zu bewerten, sondern durch den Grad ihrer Verlinkung, die Suche im World Wide Web zu einem Massenphänomen und machte den Begriff des PageRank auch außerhalb des Information Retrieval bekannt.

Ausgangspunkt ist die Beobachtung, dass im klassischen, dokumentenzentrierten Information Retrieval offene Probleme existieren, denen nicht mit traditionellen Lösungsansätzen beizukommen ist. Gleichzeitig ist aber im Umfeld der digitalen Bibliotheken eine Vielzahl an hochwertigen (Meta-)Daten vorhanden, die für das Retrieval genutzt werden können. Im Gegensatz zur Suche mit Websuchmaschinen sind hier die Inhalte kontrolliert und von einer höheren Datenqualität. Im Bereich der Informetrie existiert eine Reihe an Verfahren zur Analyse solcher Metadaten; eine konkrete Nutzung dieser hohen Informationsdichte für das Dokumentenretrieval allerdings bleibt meist aus. Der Fokus dieses Beitrags liegt deshalb auf der Frage, wie die beiden Bereiche Informetrie und IR angenähert und informetrische Verfahren zum praktischen Einsatz im IR gebracht werden können.

Der folgende Artikel soll daher eine Übersicht über den aktuellen Methodenstand im Bereich der digitalen Bibliotheken, des Information Retrieval und der Informetrie darstellen, um darauf die angedachten Lösungen aufzusetzen. Der Fokus liegt auf den Nutzerbedürfnissen hinsichtlich des Rankings von Dokumenten und dem methodischen Vorgehen im Information Retrieval. Hierzu werden die einschlägigen Verfahren und Evaluationsmöglichkeiten beschrieben sowie auf weiterführende Literatur verwiesen. Analog werden die zentralen Modelle der Informetrie vorgestellt und die Verbindungen zwischen den beiden Disziplinen aufgezeigt. Der vorliegende Text, bei dem es sich um eine gekürzte und überarbeitete Fassung des ersten Teils der Dissertationsschrift von Schaer (2013b) handelt, ist daher komplementär zum einleitenden Text in diesem Focus (Schaer 2013a) zu sehen, in dem das grundsätzliche Problem und der Gedanke des Zusammenspiels von Information Retrieval und Informetrie beschrieben wird.

2. Digitale Bibliotheken

Im Laufe der letzten Dekade hat sich sowohl beim wissenschaftlichen Nachwuchs als auch bei etablierten Wissenschaftlern eine große Veränderung im Informationsverhalten vollzogen. So ist beispielsweise bei der Suche nach wissenschaftlicher Literatur eine sogenannte „Now or Never"-Mentalität zu beobachten. Dies mag mit elektronischen Zugängen zu wissenschaftlichen Dokumenten in Form von Websuchmaschinen oder digitalen Bibliotheken – zusätzlich zu den etablierten physischen Bibliotheken an Universitäten oder anderen Forschungseinrichtungen – zu tun haben. Im folgenden Kapitel werden sowohl der typische Aufbau und die Charakteristika digitaler Bibliotheken, als auch der aktuelle Stand der Forschung zum Nutzungsverhalten dieser Systeme beschrieben. Sowohl aus der Analyse der Stärken und Schwächen der Systeme als auch den Bedürfnissen der Nutzer lassen sich im weiteren Verlauf der Arbeit die konkreten Anforderungen und Forschungsfragen ableiten.

2.1 Aufbau, Charakteristika und Begrifflichkeit

Eine klassische Definition digitaler Bibliotheken geht auf die Digital Library Federation zurück. Sie ist von 1998, sehr allgemein gehalten und spiegelt das Definitionsproblem der frühen DL-Community wieder, die im späteren Verlauf als zu beliebig angeprangert wurde. In einer späteren Definition von William Arms ist diese bereits knapper und präziser formuliert:

> [...] a digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network (Arms 2000, 2).

Weiterhin unterstreicht der Autor seine eigene Aussage: „A key part of this definition is that the information is managed“. Das ausschlaggebende Attribut ist für Arms, dass eine digitale Kollektion (intellektuell) verwaltet und gepflegt wird. Dies würde einer reinen Internetsuchmaschine die Bezeichnung als digitale Bibliothek absprechen – ihr Index ist nicht im klassischen Sinne strukturiert bzw. wird nicht verwaltet. Saracevic (2009) nennt daher auch andere virtuelle Bibliotheksangebote als Beispiele, wie die rein digital existierende WWW Virtual Library,[1] die von Tim Berners-Lee ins Leben gerufen wurde, sowie die seit 1985 bestehende themenorientierte Perseus Digital Library[2] – eine Kollektion von historischen Texten und Bildern – welche die Zeitspanne zwischen Antike und dem 19. Jahrhundert abdeckt. Ein anderes Beispiel aus der Informatik und angrenzenden Disziplinen ist die ACM Digital Library.

Fox, Goncalves und Shen (2012) grenzen digitale Bibliotheken klar von Fachdatenbanken und dem Web ab, da beides Pole eines Kontinuums seien. Auch hier wird die Struktur als Unterscheidungsmerkmal angesehen:

> One dichotomy often posed about DLs is Managed vs. Comprehensive. Thus, a library is managed while the WWW is unmanaged (but closer to being comprehensive). [...] we generally use the term structure. We argue that DLs must be organized, thus having a moderate degree of structure (Fox, Goncalves und Shen 2012, 6).

Die Verwendung des Begriffs digitale Bibliothek geht wahrscheinlich auf ein Arbeitspapier von Kahn und Cerf (1988) zurück, in dem eine nationale Infrastruktur für ein sogenanntes Digital Library System skizziert wurde. Andere Begriffe, wie elektronische oder virtuelle Bibliothek, wurden lange Zeit synonym verwendet, später dann aber in ihren Bedeutungen voneinander abgesetzt. Auch heute sind unterschiedliche Begriffe in Gebrauch, die gleichbedeutend eingesetzt werden, so z.B. digitales Archiv, Repositorium oder digitales Fachportal. Allgemein kann man die folgenden Ausprägungen in Referenzdatenbanken (reference databases) und Quelldatenbanken (source databases) unterscheiden (Chowdhury 2010, 17), wobei die Referenzdatenbanken in drei Unterkategorien aufzuteilen sind:

- Bibliografische Datenbanken, die Literaturnachweise, teilweise mit Zitationen und erweiterten bibliografischen Metadaten wie einem Abstract, enthalten.
- Bibliothekskataloge, die physische Katalogsysteme von Bibliotheken nachbilden oder mehrere Bibliothekskataloge zusammenfassen. Diese Systeme enthalten meist nur sehr wenige zusätzliche Metadaten.
- Verweisdatenbanken, die Verweise auf Referenzen zu Informationen wie Name, Adresse oder Forschungsbereiche einer Person, Institution oder einem Informationssystem enthalten.

[1] <http://vlib.org/> (Zugegriffen: 12. Oktober 2013).
[2] <http://www.perseus.tufts.edu> (Zugegriffen: 12. Oktober 2013).

In Referenzdatenbanken werden nur Verweise auf die eigentlichen Informationsquellen verwaltet. Dies können Monografien, Zeitschriftenaufsätze, Web-Ressourcen, Personenverzeichnisse und vieles mehr sein. Dem entgegengesetzt existieren die Quelldatenbanken, welche den Nutzer nicht zu einer weiteren Informationsquelle weiterleiten (wie die Referenzdatenbank), da sie selbst die benötigten Informationen enthalten. Chowdhury unterteilt diese abhängig von ihrem Inhalt in:

- Numerische Datenbanken, die numerische Datensätze unterschiedlichster Form beinhalten, so z.B. Statistiken oder Umfragedatensätze.
- Volltext-Datenbanken, die Volltexte von Dokumenten, gleich welcher Form, enthalten, auch Volltext-Repositorium genannt.
- Multimedia-Datenbanken, in denen Texte, Bilder, Audio- und Video-Informationen gespeichert sind.

Darüber hinaus sind weitere fachspezifische Datenbanktypen wie Fachdatenbanken, virtuelle Fachbibliotheken oder Fachportale bekannt, die meist eine Mischform der o.g. Referenz- und Quelldatenbanken sind. Diese entstanden aus den Strukturen der Informationsversorgung durch Fachinformationszentren, kommerziellen Fachinformationsanbietern und Bibliotheken.

Allen genannten Bezeichnungen ist gemein, dass die Bezeichnungen in unterschiedlichen Fachkontexten und Disziplinen sehr heterogen verwendet werden. Dies führt in einem interdisziplinären Umfeld zu einer weitgehenden Rückbesinnung auf den Begriff der digitalen Bibliothek. Diese ist nicht immer korrekt, da sich ein Repositorium natürlich von einem Fachportal in Aufbau und Inhalt unterscheidet. Allerdings wird die digitale Bibliothek als begriffliche Klammer bzw. Oberbegriff für die vielen unterschiedlichen und teilweise spezifischeren Bezeichnungen verwendet. Es ist davon auszugehen, dass die gegenwärtig fließende Bedeutung einem weiteren Wandel unterliegt und eine „Ausweitung des Begriffs auf digitalen Content gleich welcher Art“ (Seadle 2009, 216) zu erwarten ist.

2.2 Beispielsysteme

Im folgenden Abschnitt werden einige aktuelle DL-Systeme für wissenschaftliche Literatur beispielhaft vorgestellt. Die Beispiele sind grob in vier Untergruppen zusammengefasst: (1) kommerzielle Systeme und Kataloge, (2) öffentliche Systeme und Metakataloge, (3) Open-Access-Repositorien und (4) Suchmaschinen- und Web-Crawl-Systeme. Eine tabellarische Übersicht über die jeweiligen Systeme, ihre Betreiberorganisationen, die Anzahl der verfügbaren Nachweise und Volltexte sowie des inhaltlichen Fokus ist der Tab. 1 zu entnehmen. Die Auswahl enthält sowohl nationale als auch internationale Beispielsysteme, kann aber keinen Anspruch auf Vollständigkeit erheben. Die typischen System- und Anbietergruppen sind jedoch enthalten.

2.2.1 Kommerzielle Systeme und Verlagskataloge

ScienceDirect, SpringerLink, ACM Digital Library und das Web of Science sind typische Vertreter kommerzieller Systeme. Hinter allen Systemen stehen Unternehmen bzw. Verlage, die gegen ein entsprechendes Entgelt Nutzern Zugriff auf ihre Systeme gewähren. Bei den ersten drei genannten Systemen ist der Zugang zu den erfassten Literaturnachweisen und einem Großteil der Metadaten kostenfrei möglich. Für zahlende Kunden wird darüber hinaus ein erweiterter Zugriff u.a. auf Volltexte, Zitationsdaten und weitere Mehrwertdienste gewährt. Diese Systeme sind also sowohl Nachweis- als auch Volltextdatenbanken, die sich zu einem großen Teil aus den Kataloginformationen der jeweiligen Betreiber speisen. Darüber hinaus werden durch Kooperationsverträge mit anderen Anbietern die Nachweissysteme mit externen Informationen erweitert. Im Gegensatz zu den drei vorgenannten Systemen ist das Web of Science eine reine Nachweisdatenbank, die allerdings einen starken Fokus auf eine breite fachliche Abdeckung und eine Erfassung von Zitationen und Referenzen legt.

Die Qualität der angebotenen Systeme ist gemeinhin sehr hoch, da die hinterlegten Nachweisinformationen entweder direkt aus den Katalogen der Verlage stammen, eingekauft oder mit großem Ressourcenaufwand selbst erstellt werden. Der kommerzielle Charakter der Systeme erlaubt eine wissenschaftliche Auswertung meist nur über Kooperationsverträge bzw. gegen ein entsprechendes Entgelt.

Tab. 1: Beispiele für digitale Bibliotheken

Name	Betreiber	Art	NW	VT	Domäne
BASE	Universität Bielefeld	S	51 673	51 673	Interdisziplinär
Web of Science	Thomson Reuters	K	31 093	--	Interdisziplinär
ScienceDirect	Elsevier	K	11 991	11 991	Interdisziplinär
Google Scholar°	Google	S	10 900	5 410	Interdisziplinär
SpringerLink	Springer	K	7 949	7 949	Interdisziplinär
Sowiport	GESIS	P	7 444	600	Sozialwissenschaften
Europeana Libraries°	Europeana Foundation	P	5 000	5 000	Interdisziplinär
DBLP	Uni Trier	P	2 412	--	Informatik
ACM Digital Library	ACM	K	2 165	392	Informatik
CiteseerX	Pennsylvania SU	S	1 472	1 472	Informatik
RePEc	dezentral organisiert	P	1 459	1 320	Wirtschaft
arXiv	Cornell University	O	881	881	Physik, Informatik, Mathematik
FIS Bildung	DIPF	P	819	55	Pädagogik
EconStor	ZBW	O	63	63	Wirtschaft
SSOAR	GESIS	O	26	26	Sozialwissenschaften
pedocs	DIPF	O	5	5	Pädagogik

Es wird unterschieden zwischen kommerziellen Systemen und Katalogen (K), öffentlichen Systemen und Metakatalogen (P), Open-Access-Repositorien (O) sowie Suchmaschinen- und Web-Crawl-Systemen (S). Die Größe der Systeme ist in Tausend angegeben jeweils für die Anzahl der Nachweise (NW) und der Volltexte (VT). Die Daten beziehen sich auf den Stand Oktober 2013. Systeme, deren Größe nur geschätzt werden konnte, sind durch ° markiert.

2.2.2 Öffentliche Systeme und Metakataloge

Eine Alternative zu den kommerziellen Systemen stellen die mit öffentlichen Mitteln geförderten Systeme Sowiport, Europeana Libraries, RePEc, DBLP oder FIS Bildung dar. Auch hier liegt der Fokus auf der Erfassung von Literaturnachweisen, obwohl auch vermehrt Volltexte bzw. Verlinkungen zu Volltexten in die Systeme aufgenommen werden. Die vier Systeme Sowiport, RePEc, DBLP und FIS Bildung besitzen im Gegensatz zu Europeana Libraries einen engen fachlichen Fokus. Für Sowiport sind dies die Sozialwissenschaften, für RePEc die Wirtschaftswissenschaften, für DBLP die Informatik und für FIS Bildung die Pädagogik. Europeana Libraries dient als Sammelstelle für wissenschaftliche Literatur und kooperiert mit wissenschaftlichen Bibliotheken in ganz Europa, allerdings ohne einen vergleichbaren inhaltlichen Fokus wie die zuvor genannten Systeme.

Ähnlich zu den kommerziellen Anbietern stehen die betreibenden Organisationen für eine hohe Qualität der angebotenen Daten ein. Über Kooperationsverträge oder Lizenzen werden auch teilweise Inhalte kommerzieller Anbieter (z.B. im Rahmen der Nationallizenzen der Deutschen Forschungsgemeinschaft, DFG) in die Systeme eingebunden und der Öffentlichkeit zur Verfügung gestellt. Dies ermöglicht auch vielfach den kostenfreien Zugriff auf die vorgehaltenen Daten zum Zwecke der wissenschaftlichen Untersuchung.

2.2.3 Open-Access-Repositorien

Einen klaren Fokus auf die Akquise von Volltexten legen sogenannte Open-Access-Repositorien. Diese Dokumentenserver werden meist von Hochschulen oder Forschungseinrichtungen betrieben und lassen sich in institutionelle und disziplinäre Angebote unterteilen. Institutionelle Angebote sammeln vorwiegend die Publikationen der Angestellten der eigenen Institution und stellen diese frei zugänglich über das Repositorium zur Verfügung. Disziplinäre Repositorien sammeln nicht nur mit dem Fokus auf eine Institution, sondern für einen bestimmten Fachbereich. Die vier in Tab. 1 vorgestellten Systeme arXiv, EconStor, SSOAR und pedocs sind ausnahmslos disziplinäre Repositorien, jeweils für die Fachgebiete Physik und Informatik, Wirtschaftswissenschaften, Sozialwissenschaften und Pädagogik.

Die Volltexte werden zu einem großen Teil auf den Systemen selbst gespeichert und üblicherweise nur rudimentär erfasst. Dies liegt vor allem an der häufig durchgeführten Selbsterschließung der Nachweise durch die Autoren selbst. Im Open-Access-Umfeld hat sich die Erfassung mit Dublin Core und der Klassifizierung nach DDC als kleinster gemeinsamer Nenner etabliert. Da alle in den Repositorien gesammelten Inhalte kostenfrei (meist auch unter einer sehr offenen Lizenz wie z.B. den Creative Commons) angeboten werden, ist eine wissenschaftliche Nachnutzung der Nachweise und der Volltexte problemlos möglich. Mit OAI-PMH (Open Archive Initiative Protocol for Metadata

Harvesting) existiert sogar ein eigenes Austauschprotokoll, um die Inhalte von Open-Access-Repositorien zu sammeln. Dieser Vorgang wird auch als Harvesting bezeichnet.

2.2.4 Suchmaschinen- und Web-Crawl-Systeme

Einen übergreifenden Zugang zu den zuvor genannte Systemen möchten Suchmaschinen wie z.B. BASE, Google Scholar und CiteseerX bereitstellen. Im Gegensatz zu den bisher vorgestellten halten diese Systeme die Nachweis- bzw. Volltextdaten nicht selbst vor, sondern pflegen nur einen Index, in dem nach den Informationen gesucht werden kann. Die eigentlichen Nachweise und Volltexte sind über Harvestingvorgänge (BASE) oder Web-Crawling (Google Scholar und CiteseerX) gesammelt worden. Systembedingt haben Suchmaschinen daher oft ein Dublettenproblem. Während die beiden Systeme BASE und CiteseerX öffentlich finanziert und somit auch die Inhalte frei verfügbar sind, ist Google als kommerzieller Anbieter nur schwer zu untersuchen. So musste beispielsweise die Größe von Google Scholar nach einem Verfahren von Jacsó (2008) geschätzt werden, da es als einziges System keine genaue Größenangaben enthält. Die Datenqualität ist stark abhängig von den jeweiligen Datenquellen und ist, da das primäre Interesse nur auf dem Füllen eines Suchindex‘ liegt, nicht mit denen von kommerziellen oder öffentlichen Nachweisverzeichnissen zu vergleichen.

2.3 Metadaten in digitalen Bibliotheken als Schlüssel zur Suche

In allen bisher beschriebenen Varianten von Informationssystemen wird den sogenannten Metadaten eine besondere Bedeutung beigemessen. Metadaten, die primär zunächst „Daten über Daten“ sind, werden eingesetzt, um Daten oder Objekte strukturiert zu beschreiben. In frühen digitalen Bibliotheken und den zuvor genannten Referenz- und Quelldatenbanken war die Suche über diese Metadaten die einzige Möglichkeit, die hinterlegten Daten und Dokumente zu finden. Die Erschließungsqualität der Metadaten hatte direkte Auswirkung auf die Suchmöglichkeiten innerhalb der Referenzobjekte, die durch die Metadaten beschrieben wurden (Voß 2007).

Dennoch war und ist eine direkte Suche innerhalb der Datenbestände in vielen Systemen nicht möglich. In großen digitalen Bibliotheken, wie z.B. der Europeana, die neben textuellen Dokumenten auch Bild-, Audio- und Videomaterial enthält, ist die Suche über die Metadaten die einzige Möglichkeit auf die Datenbestände zuzugreifen. Obwohl die direkte Suche in solchen nichttextuellen (in diesem Kontext multimedialen) Datenbeständen große Fortschritte macht, ist diese Art der Suche bisher noch Spezialsystemen und der Forschung vorbehalten – im Moment dominiert die text-orientierte Suche in Metadaten (Neal 2012).

Auch in Systemen, die Volltexte durchsuchbar machen, sind Metadaten von Bedeutung. Eine Volltextsuche ist beispielsweise nicht in der Lage, zwischen dem Autor eines Textes und der Erwähnung eines Namens in den Zitationen zu unterscheiden. Eine entsprechende Auszeichnung der Metadaten nach klar definierten Regeln erlaubt eine gezielte Suche u.a. nach Personen, Institutionen oder Publikationsjahren. In bibliografischen Informationssystemen gibt es daher mehrere Standards zur Beschreibung von Publikationen und Dokumenten unterschiedlichster Art. Diese sorgen für eine Normierung der Erschließungsvorgänge und erlauben eine Grundannahme über die Qualität der Erschließung. Beispiele hierfür sind deutsche Normen wie DIN 1502 oder DIN 1505-2, die Regeln zur Vereinheitlichung von Zeitschriftennamen oder Zitierungen beinhalten. Für die Entwicklung und Anwendung von Thesauri gibt es die DIN 1463-1 und das internationale Äquivalent ISO 2788, die unterschiedliche Relationsarten innerhalb der Thesauri oder Abkürzungen beschreiben. Seitens der Deutschen Nationalbibliothek wurde 2012 die Vielzahl an verschiedenen Normdateien zur „Gemeinsamen Normdatei" zusammengefasst.

Im folgenden Abschnitt werden die Erschließungsregeln für die in diesem Beitrag betrachteten Dokumentarten Monografien, Sammelwerke und Zeitschriftenaufsätze kurz beschrieben, da auf die jeweiligen Eigenschaften später bei der Entwicklung der Retrievalmodelle zurückgegriffen wird.

2.3.1 Zeitschriften

Zur eindeutigen Identifizierung von Zeitschriften wurde die Internationale Standardnummer für fortlaufende Sammelwerke (International Standard Serial Number, ISSN) entwickelt, die von der ISO unter der Nummer 3297 zertifiziert wurde. Die Vergabe einer international verbindlichen und weltweit einmaligen Nummer zur Identifikation von Zeitschriften ist notwendig, da diese über die oft lange Lebenszeit einer Zeitschrift unter verschiedenen Namen und Schreibweisen bekannt sein können. Die Zeitschrift „Schmollers Jahrbuch – Journal of Applied Social Science Studies" (ISSN 1439-121X) hat seit ihrer Gründung 1871 als „Jahrbuch für Gesetzgebung, Verwaltung und Rechtspflege des Deutschen Reiches" unterschiedliche Namen getragen.

Das Problem der unterschiedlichen Namensgebung über die Lebensdauer wird durch die Vergabe von ISSN nur indirekt gelöst, da entsprechend der Richtlinien des ISSN International Centre die ISSN bei jeder Namensänderung und bei Änderung der Publikationsform von Print- auf Onlinepublikation ebenfalls geändert werden muss. Dies ermöglicht es aber zumindest, an einer zentralen Stelle die verzeichneten Änderungen nachzuvollziehen. Das ISSN Portal[3] ermöglicht nach vorheriger kostenpflichtiger Registrierung einen zentralen Zugriff auf die ISSN-Bestände über die Webseite des ISSN International Cent-

[3] <http://portal.issn.org> (Zugegriffen: 12. Oktober 2013).

re. Somit kann die ISSN trotz dieser Einschränkung als wichtiges Werkzeug zur klaren Unterscheidung von Zeitschriften gesehen werden, da durch ihre Verwendung die Probleme der Ansetzungsformen von Zeitschriftennamen, Sonderausgaben, mehrsprachigen Titeln oder simpler Rechtschreibfehler vermieden werden.

Die achtstellige ISSN wird aus den Ziffern 0-9 gebildet und in zwei vierstellige Blöcke aufgeteilt. Die letzte Stelle ist dabei eine Prüfziffer, die sich durch die Differenz von 11 zum Modulo 11 der Quersumme berechnet. Die Quersumme wird zuvor von der ersten bis zur letzten Stelle von 8 bis 2 absteigend gewichtet. Anstelle des Wertes 10 wird das Zeichen X angegeben. Neben der Prüfziffer ist aber kein weiteres strukturierendes Element enthalten. Weder gibt es einen Code, der Verlag oder Herkunftsland beschreibt, wie dies bei der Internationalen Standardbuchnummer der Fall ist, noch ist eine Information über die Periodizität, d.h. die Erscheinungshäufigkeit, enthalten. Die ISSN wird international dezentral vergeben. In Deutschland ist die Deutsche Nationalbibliothek für die Zuteilung der ISSN zuständig. Insgesamt sind bis zum Jahr 2010 mehr als 1,55 Millionen Nummern vergeben worden.

2.3.2 Monografien

Ähnlich zu den ISSN für Periodika gibt es die Internationale Standardbuchnummer (International Standard Book Number, ISBN), die strukturelle Informationen beinhaltet. Sie ist als DIN 1462 bzw. als ISO 2108 genormt. Die ISBN besteht aus dreizehn Stellen und ist in fünf Teile aufgeteilt: Präfix (dreistellig, wird international vergeben), Gruppennummer für nationale, geographische Sprach- oder ähnliche Gruppen (ein- bis fünfstellig, Vergabe durch nationale ISBN-Agentur; die Ziffer 3 steht für deutschsprachige Länder), Verlagsnummer für den einzelnen Verlag innerhalb einer Gruppe (zwei- bis siebenstellig, Vergabe durch nationale ISBN-Agentur), Titelnummer für das einzelne Buch des in Teil 3 bezeichneten Verlages (ein- bis sechsstellig, Vergabe durch Verlage aus einem festen Nummernkontingent), Prüfziffer (einstellig). Durch die entsprechende Aufteilung ist folglich in der ISBN sowohl das Herkunftsland – und damit indirekt die Sprache – als auch der herausgebende Verlag kodiert. Ein großer Verlag, dem eine dreistellige Verlagsnummer zugeordnet wurde, kann bis zu 100.000 Titelnummern vergeben. Bis Ende 2006 waren ISBN nur zehnstellig und das dreistellige Präfix fehlte. Eine zehnstellige ISBN kann aber im Falle von Büchern mit dem konstanten Präfix 978 versehen werden und ist damit eine korrekte dreizehnstellige ISBN. Wie auch bei der ISSN muss bei einer Änderung der Monographie, z.B. einer Neuauflage, auch die ISBN angepasst werden.

Durch die vorgegebene Struktur der ISBN kann diese analysiert werden.[4] Die dreizehnstellige ISBN des Buchs „Introduction to modern information retrieval“ mit der ISBN 978-1-85604-694-7 erlaubt z.B. folgende Rückschlüsse auf das Dokument: Die ISBN-10 lautet 1-85604-694-X, beide ISBN-Varianten sind in Gebrauch, sie gehören der Gruppe „Englischsprachige Gebiete“ an und der Verlag „Facet Publishing“ stammt aus dem Vereinigten Königreich oder Irland.

2.3.3 Autorennamen

Die Namen von Autoren oder Herausgebern (im Englischen auch allgemeiner als „creator“ bezeichnet) sind ein weiterer elementarer Bestandteil von in digitalen Bibliotheken erfassten Metadaten. Vielfach werden Autorennamen in einfacher Textrepräsentation erfasst. Dies führt aber, wie auch schon bei Zeitschriften oder Monografien beschrieben, zu gewissen Ungenauigkeiten. Am Beispiel des Autors ist dies anschaulich zu zeigen, da er selbst in verschiedenen Ansetzungsformen in unterschiedlichen Datenbanken verzeichnet ist: Sei es, dass der Nachname mit einem Umlaut oder ohne verwendet wird (Schaer/Schar/Schär), der Vorname unterschiedlich abgekürzt wird (Philipp, Ph., P. oder einfach P), oder Vor- und Nachname nicht richtig auseinandergehalten werden. Natürlich sind auch beliebige Mischungen dieser Fehler möglich. Neben den Ansetzungsformen sind mehrdeutige Namen ein Problem, wie van Noorden[5] anschaulich aufzeigt: „Most-published researchers in 2011? Wang Y, Zhang Y, Liu Y, Wang J, Li Y, Wang L, Li J, Zhang J, Zhang L, Kim JH“.

Scharnhorst und Garfield (2010) erkennen ebenfalls das Identifikationsproblem und beschreiben drei unterschiedliche Lösungsansätze mit denen dem Problem entgegen gearbeitet werden soll: (1) mittels eindeutiger Personenidentifikatoren, (2) einheitlicher Erschließungsregeln und (3) automatischer Extraktion und Zusammenführung heterogener Datenbestände.

Sie geben eine Übersicht über die unterschiedlichen Systeme zur Vergabe eines eindeutigen digitalen Identifikators für Autoren und unterscheiden dabei zwischen kommerziellen und öffentlich finanzierten Systemen, die nebeneinander, miteinander, und auch gegeneinander arbeiten. Kommerzielle Verlage wie z.B. Thomson Reuters oder Elsevier haben eigene Systeme zur Autorenidentifikation eingeführt, wie z.B. die ResearcherID[6]. Demgegenüber stehen öffentliche Systeme, wie der Digital Author Identifier (DAI) der niederländischen SURF Foundation oder der Personennamendatei (PND) der deutschen Nationalbibliothek (DNB). In diesen Systemen werden sowohl normierende

[4] <https://toolserver.org/isbn/IsbnCheckAndFormat> (Zugegriffen: 12. Oktober 2013).
[5] <http://twitter.com/#!/Richvn/status/144414256610283520> (Zugegriffen: 12. Oktober 2013).
[6] <http://www.researcherid.com> (Zugegriffen: 12. Oktober 2013).

Regeln für die Erschließung, als auch eindeutige Identifikatoren vergeben. Im Falle der DNB ist dies die sogenannte PND-Nummer.

Außer den Namen werden auch Pseudonyme, Beruf und weitere ergänzende Daten erfasst. Neben den kommerziellen und öffentlich-finanzierten Anbietern gibt es auch Bestrebungen, die aus der wissenschaftlichen Gemeinschaft selbst stammen. Beispiele hierfür sind das Identifikatorenprogramm des arXiv[7] oder das Authorclaim-System[8] des RePEc-Initiators Thomas Krichel. Das System ORCID[9] ist eine neue Entwicklung, das zum Ziel hat, vorhandene Lösungen wie die ResearcherID von Thomson Reuters, die Autorenprofile von RePEc, Scholar Universe und Scopus in einer einheitlichen technologischen Basis zusammenzubringen.

2.3.4 Kontrollierte Schlagwörter und Thesauri

Ein Thesaurus liefert ein sogenanntes kontrolliertes Vokabular, das ermöglicht Dokumente zu beschreiben. Diese terminologische bzw. begriffliche Kontrolle ist eine der größten Stärken von Thesauri, da sie durch die Kombination der kontrollierten Begriffe (auch Deskriptoren genannt) eine große Ausdrucksmächtigkeit und -klarheit schaffen. Dies kann insbesondere für den Retrievalprozess genutzt werden, da hierbei das eingeschränkte Vokabular eine gezieltere Suche ermöglicht – vorausgesetzt, der Suchende ist in der Lage, das Vokabular des Thesaurus korrekt zu verwenden.

Ein konkretes Beispiel für einen Thesaurus ist der Thesaurus Sozialwissenschaften, der von GESIS entwickelt und gepflegt wird. Er wird hauptsächlich für die sozialwissenschaftliche Literaturdatenbank SOLIS sowie die Forschungsdatenbank SOFIS verwendet und liegt in insgesamt drei Sprachen vor (deutsch, englisch und russisch). Die deutsche Ausgabe enthält ca. 8.000 Deskriptoren und 4.000 Nicht-Deskriptoren, die fachlich die Disziplinen der Sozialwissenschaften abdecken. Der Thesaurus ist online verfügbar[10] und kann auch als SKOS-Version abgefragt werden.

2.4 Nutzerverhalten und -bedürfnisse

In Bezug auf digitale Bibliotheken stehen aktuelle Systeme vor der Aufgabe, Nutzern qualitativ hochwertige (Fach-)Information auf schnellem Wege und mit einfacher Bedienbarkeit bereitzustellen. Aktuelle Studien, wie z.B. eine JISC-Studie von Wong et al. (2010), zeigen, dass durch die Angebote moderner Websuchmaschinen die Erwartungshaltung der Nutzer schneller wächst, als ihre Fähigkeit, mit den Systemen umzugehen:

---

[7] <http://arxiv.org/help/author_identifiers> (Zugegriffen: 12. Oktober 2013).
[8] <http://authorclaim.org> (Zugegriffen: 12. Oktober 2013).
[9] <http://orcid.org/> (Zugegriffen: 12. Oktober 2013).
[10] <http://www.gesis.org/sowiport/nc/suche/thesaurus.html> (Zugegriffen: 12. Oktober 2013).

> For example, at a simplistic level, many of the participants do not understand how to assess the quality of materials they find. Google or Google Scholar have lower thresholds of information technology literacy, and are considered their '... friends' because of the apparent higher yield or success rate (Wong u.a. 2010, 5).

Wie Rick Anderson (2011) beschreibt, befinden sich die wissenschaftlichen Bibliotheken aktuell in einer „Crisis in Research Librarianship“, womit der Zustand, dass in reinen Zahlen immer weniger Nutzer die Fachdienstleistung einer wissenschaftlichen Präsenzbibliothek in Anspruch nehmen, gemeint ist. So habe sich in den letzten 15 Jahren, nach den offiziellen Zahlen der Association of Research Library (ARL), die Zahl der Geschäftsgänge (reference transactions), die sich innerhalb einer der erfassten Bibliotheken ereignet haben, mehr als halbiert. Angesichts der steigenden Zahl an Studenten pro Jahrgang sei der Rückgang noch drastischer: Waren 1995 noch 10,1 Geschäftsgänge pro eingeschriebenem Student zu verzeichnen, waren dies 2009 nur noch 3,6. Für Anderson sind diese Zahlen ein Indiz für die These, dass Nutzer wissenschaftlicher Bibliotheken (in diesem Falle Studenten), ihre Informationsbedürfnisse zunehmend außerhalb der traditionellen Institutionen befriedigen – zum einen, weil es durch die Vielzahl an elektronischen Zugriffsmöglichkeiten umfangreiche Alternativen gibt und zum anderen, weil sich die Nutzer zunehmend als fähig betrachten, Informationen selbst zu recherchieren und keine Unterstützung durch Bibliothekare mehr in Anspruch nehmen. Weitergehend kann gesagt werden, dass Präsenzbibliotheken bei der Informationssuche keine wichtige Rolle mehr spielen, wohl aber bei der Beschaffung oder der Lizenzierung von Datenbanken.

Diese Thesen sind angesichts steigender Zugriffszahlen in digitalen Bibliotheken, Fachportalen und anderen Online-Angeboten schlüssig. Allerdings zeigen verschiedene Studien, dass zwar quantitativ eine Bewegung weg von den Präsenzbibliotheken und dortiger Beratungen zu sehen ist, diese Entwicklung aus Sicht der Nutzer jedoch sehr ambivalent ausfällt. So werden entfallene Fußwege und eine größere Datenbasis als positiv bewertet, doch viele Nutzer der elektronischen Angebote sind bei der eigenständigen Suche schlichtweg überfordert, trotz spezieller Hilfestellungen wie Schritt-für-Schritt Anleitungen oder Fibeln, die speziell für die wissenschaftliche Online-Recherche ausgelegt sind. Zu gleichen Ergebnissen kamen auch andere Studien: Nutzer sind bei der Suche häufig überfordert und werden durch DL-Systeme nicht genügend unterstützt.

In anderen Studien wird der „Information Overload“ als eines der Hauptprobleme bei der Suche von wissenschaftlicher Literatur herausgearbeitet, wobei zwischen einer nachfrageinduzierten und einer angebotsinduzierten Ausprägung des Problems unterschieden wird:

> Im Fall des nachfrageinduzierten Overload-Problems liegt die Ursache vor allem in der mangelnden Informationskompetenz (information literacy) der Benutzenden [...]. Im Fall des angebotsinduzierten Problems übersteigt das An-

> gebot, trotz ausreichender Kenntnisse und Fähigkeiten bei der Recherche, das Maß der Informationsmenge, die für die jeweiligen Wissenschaftler konsumierbar ist (te Boekhorst, Kayß und Poll 2003, 4-5).

Die Autoren der Studie weisen auf zwei Strategien der Informationssuchenden hin, um das Overload-Problem zu beherrschen. Zum einen vertrauen Wissenschaftler primär auf ihre persönlichen Informationsnetzwerke (invisible colleges), die sie mit aktueller und relevanter Literatur versorgen, zum anderen wird vermehrt nur die Art von elektronischer Information wahrgenommen, die aktuell verfügbar ist („Verfügbarkeit hat Priorität", ebenda, 8). Man spricht hierbei auch von einer „Now or Never-Mentalität" (Sühl-Strohmenger 2008, 58).

2.5 Relevanzbeurteilungen durch Nutzer

Für die Zentralbibliothek der Wirtschaftswissenschaften (ZBW) haben Siegfried und Flieger (2011) in einer Nutzerstudie mit 160 Studenten und Forschern (wissenschaftliche Mitarbeiter, Doktoranden, Juniorprofessoren, Professoren oder Forschende) die zuvor aufgestellten Thesen zum Nutzungsverhalten in DL-Systemen untersucht. Tatsächlich berichteten die Nutzer über teilweise erhebliche Probleme bei der Recherche. So werden u.a. Probleme bei der Beurteilung der wissenschaftlichen Qualität eines Suchtreffers, bei der Einordnung der Relevanz für das eigene Forschungsvorhaben und bei der Suche nach einem passenden Schlagwort angeführt. Ähnliche Ergebnisse präsentiert auch eine OCLC-Studie (Calhoun u.a. 2009, 11), die einen eindringlichen Appell der Nutzer zusammenfasst: „Search results must be relevant and the relevance must be obvious."

Gleichzeitig zeigt die ZBW-Studie sehr gut, wie und nach welchen Kriterien die Teilnehmer der Studie die Relevanz von Suchergebnissen sortieren und deren Einschlägigkeit beurteilen. Die drei wichtigsten Kriterien sind zunächst die Aktualität der Suchergebnisse (gemeint ist das Publikationsdatum), gefolgt vom Renommee der Zeitschrift und des Autors. Weiterhin werden die Häufigkeit der Zitation und die Anzahl der Downloads als Qualitäts- und damit Relevanzindikator herangezogen (Siegfried und Flieger 2011, 6-7). Wie aber schon zuvor beschrieben, sind diese Relevanzkriterien für 53% der Forschenden nur schwer nachzuvollziehen und einzuordnen.

Weiterhin greifen Forschende bei der Einordnung nach Relevanz gerne auf „nahestehende" Experten zurück, sowohl bei der Suche als auch. Die meist dem sogenannten Web 2.0 zugeordnete Funktionalität der Bewertung von Dokumenten durch Benutzer (social feedback) wird von Wissenschaftlern hingegen meist als störend wahrgenommen – ausgenommen davon sind professionelle Reviews und Expertenmeinungen (Calhoun u.a. 2009, 18).

Neben Studien, die auf Nutzerbefragungen und Interviews setzen, werden zunehmend auch Analysewerkzeuge, wie z.B. Google Analytics, eingesetzt, um Studien zum Nutzerverhalten (User Behaviour Studies) durchzuführen. So

wurde im Rahmen einer Logfile-Analyse des sozialwissenschaftlichen Portals Sowiport von Schaer et al. (2012) ermittelt, dass von der Vielzahl an erfassten Metadaten nur wenige in der direkten Suche eingesetzt werden. Bei der Analyse der 1.000 populärsten Suchanfragen (insgesamt N=129.251) zeigte sich, dass in knapp 1/3 der Fälle explizit nach einer Person bzw. deren Namen gesucht wurde; nur 1/5 der Benutzer verwendete explizit kontrollierte Schlagwörter für die Suchanfrage. Die weiteren vorhandenen Suchfelder für Publikationsjahr, Quelle, Institutions- oder Ortsnamen wurde in weniger als 1% der untersuchten Fälle verwendet. Nahezu die Hälfte der Suchanfragen waren einfache Freitextsuchen, die alle vorhandenen Metadaten abdeckten, sodass hier keine bestimmten Entitäten identifiziert werden konnten.

Das Suchverhalten der Benutzer im Portal Sowiport kann dabei als relativ typisch für die Nutzer in digitalen Bibliotheken allgemein angesehen werden: In vielen Fällen wird (ob aus Unkenntnis oder Unzufriedenheit angesichts der Alternativen) auf erweiterte Suchunterstützungsmöglichkeiten in Form spezialisierter Dienste oder Suchmasken nicht zurückgegriffen. Dies mag daran liegen, dass – entgegen des tatsächlichen Nutzerinteresses an bspw. Personen – Dienste angeboten werden, die speziell auf kontrollierte Vokabulare abzielen, wie z.B. in Form sogenannter Search Term Recommender.

Tab. 2: Ergebnis einer Logfile-Analyse des sozialwissenschaftlichen Portals Sowiport

Suchentitäten	Anzahl der Anfrage	Anteil an Gesamtanfragen
Freitextsuche (alle Felder)	58.754	45,5 %
Person(en)	40.979	31,7 %
Schlagwort	26.959	20,9 %
Titel	1.108	0,9 %
Quelle	581	0,4 %
Andere	354	0,3 %

Analysiert wurden die 1.000 populärsten Suchanfragen (insgesamt N=129 251). Aufgelistet werden die Anzahl der Fragen nach Suchentität und deren prozentualen Anteil an der Gesamtmenge der Anfragen. Die letzte Zeile fasst die Entitäten Publikationsjahr, Institution und Ort zusammen (Tabelle angelehnt an Schaer et al. 2012).

2.6 Offene Probleme im Dokumenten Retrieval

Ein essentielles Problem digitaler Bibliotheken ist die nutzerseitige Suche und das damit verbundene Ranking der Ergebnismenge. Bedingt durch die Charakteristika der Systeme, welche meist aus intellektuell verwalteten Metadaten bestehen, sind viele anspruchsvolle und komplexe Retrieval- und Rankingfunktionen nicht anwendbar. Dies ist größtenteils auf das Fehlen passender Daten zur Berechnung dieser Verfahren zurückzuführen. So scheitert z.B. eine Linkanalyse am Fehlen solcher Verknüpfungen der Metadaten untereinander, bzw. an einer ausreichenden Abdeckung der Metadaten mit den dazugehörigen Volltexten, die z.B. eine Zitationsanalyse zuließe. Weiterhin sind nicht in allen

Systemen Mechanismen zur Verwaltung von User-Profilen eingebaut, die eine Auswertung von Nutzungsverhalten und Relevance-Feedback erlauben würden.

In einer Studie von Buckley (2009) zeigte sich, dass einige seiner Testsysteme einen Großteil der vorhanden Metadaten schlichtweg ignorierten und so die Retrievalsysteme nicht darauf zurückgreifen konnten. Selbst wenn die Metadaten indexiert wurden, wurden sie meist nicht für das Retrieval verwendet. Buckley stellt die Frage, ob Potential verschenkt wurde und ob zusätzlich zu sprachzentrierten Verbesserungen wie Stemming oder einer NLP-getriebenen Anfrageerweiterung andere Eigenschaften der Dokumente hätten genutzt werden können, um das Suchergebnis zu verbessern. Buckley fasst in der Debatte die folgenden essentiellen Probleme aktueller Retrievalsysteme zusammen: (1) technische Fehler (Stemming, Tokenization etc.), (2) Überbewertung oder Fehlbewertung einzelner Begriffe oder Aspekte in der automatischen Anfragegenerierung und (3) weitere Probleme, die aus Bereichen des Natural Language Understanding (NLU) und Natural Language Processing (NLP) stammen. Neben den sprachzentrierten Problemen wurde unter anderem festgestellt, dass die in den Systemen vorhandenen Metadaten nur unzureichend für das Retrieval genutzt wurden. In der mit mehr als tausend Personenstunden und insgesamt sieben bewerteten Retrievalsystemen groß angelegten Studie kam man zu dem Schluss, dass nicht-term-basierte Sucheigenschaften das Suchergebnis erheblich verbessern können. Im Falle der Buckley-Studie war dies die Angabe der Zeitschrift Financial Times, in der die Artikel erschienen.

Wirft man einen Blick auf typische digitale Bibliotheken wie die ACM Digital Library, das CiteseerX-System, das Web of Science oder ScienceDirect von Elsevier, werden dem Benutzer nur eine recht begrenzte Anzahl an Sortierungsmöglichkeiten geboten, ein wirkliches Ranking ist meist nur in Form von Relevance-Ranking implementiert. Neben klassischem text-basierten Relevanzranking wird die Umsortierung sowohl nach nominellen Kriterien (Erscheinungsjahr) als auch alphabetischen Kriterien (Titel oder Herausgebername) angeboten.

Systeme wie z.B. arXiv verzichten vollständig auf ein Ranking der Dokumente und bieten nur ein boolesches Retrieval mit einfacher Sortierung nach Eingangsdatum oder alphabetischen Kriterien an. Im Falle von ScienceDirect kommt hingegen neben einfachen Sortierkriterien wie Erscheinungsjahr oder Downloadzahlen das bekannte TF*IDF-Ranking hinzu. Die Popularität des sogenannten Vektorraummodells für digitale Bibliotheken zeigt sich auch in der Nutzung der freien Suchmaschinentechnologie Solr, die bei Citeseer oder der OPAC-Software Blacklight[11] eingesetzt wird. Blacklight ist als freier „next-

[11] <http://projectblacklight.org/> (Zugegriffen: 12. Oktober 2013).

generation catalog“ für den Einsatz mit heterogenen Dokumentenkollektionen konzipiert worden.

Lewandowski (2009) stellt eine Übersicht über Rankingfaktoren zusammen, die in digitalen Bibliothekssystemen Verwendung finden. Er schlägt fünf unterschiedliche Kategorien dieser Faktoren vor, die genutzt werden können, um Felder in der Datenbank entsprechend der Herkunft ihrer – textuellen – Information zu gewichten. Seiner Beschreibung nach sei es „best practice“, Titelterme höher in der Rankingformel zu gewichten als Terme, die im Abstract eines Dokuments enthalten sind. Durch den rein textuellen Abgleich von Anfrage- und Dokumenttermen wird das Retrievalsystem allerdings dazu gezwungen, die zuvor sorgsam codierte Semantik zu ignorieren – eine Eigenschaft, die allen Bag-of-Words-Ansätzen gemein ist. Neben Termgewichtungen werden noch Faktoren wie Popularität, Neuigkeit oder Lokalität beschrieben. Hinzu kommen sehr domänenspezifische Eigenschaften, die nur im konkreten Anwendungsfall eines OPAC einer Präsenzbibliothek sinnvoll sind. Beispiele hierfür sind der Aufenthaltsort des Benutzers und die physischen Entsprechungen der digitalen Objekte in Form von Printmedien in einem Regal.

Andere von ihm aufgeführte Faktoren sind allerdings auch auf andere Anwendungsfelder übertragbar, wie die Dokumentengröße, Dokumententyp oder Popularitätswerte wie die Anzahl der Dokumentenansichten, Benutzerbewertungen, Zitationen etc. Speziell in modernen, Web-basierten digitalen Bibliotheken erlangt diese Art der Rankings zusehends an Zuspruch. Eine weitgehende Auswertung und Evaluation ist diesen Faktoren aber bislang noch nicht gewidmet worden, von einigen Papieren wie z.B. von Schlögl und Gorraiz (2012) abgesehen.

3. Menschliche Relevanzeinschätzung und maschinelles Ranking

Wie im vorherigen Kapitel dokumentiert, sind traditionelle digitale Bibliotheken zwar mit einer Fülle an meist gut strukturierten Metadaten gefüllt, doch bestehen klare Defizite in einem nutzergerechten Retrieval innerhalb dieser Datenbestände. Im folgenden Kapitel werden daher zunächst sowohl die traditionellen Rankingverfahren aus dem Ad-hoc-Dokumentenretrieval, als auch aktuelle Verfahren aus dem Web-Retrieval und verwandten Disziplinen aufgezeigt, um den aktuellen Stand der Technik zu dokumentieren. Zunächst werden die vier traditionell textuell-orientierten Verfahren des booleschen Modells, des Vektorraummodells, des probabilistischen Modells und der statistischen Sprachmodelle vorgestellt. Im nächsten Abschnitt folgen eine Reihe primär nicht-textueller Rankingverfahren, die z.B. wie der PageRank auf Link-Analyse oder wie die Autorenzentralität auf einer Analyse der Ko-Autorenschaften basieren. Das Kapitel schließt mit einer Gegenüberstellung

beider Ansätze des automatisierten und interaktiven IR unter Berücksichtigung der historischen Debatte zwischen Agentensystemen und direkter Manipulation.

In diesem Beitrag werden sogenannte textuelle und nicht-textuelle Verfahren und Konzepte unterschieden. Während in der Literatur nicht-textuell häufig mit multimedial gleichgesetzt wird, sind damit im Rahmen dieses Artikels Eigenschaften gemeint, die keinen textuellen Bezug haben oder zumindest nicht textuell behandelt werden. Metzler (2011, 5) führt einige Beispiele für solche nicht-textuellen Eigenschaften auf, u.a. den PageRank, das Zählen von Inlinks einer Netzwerkanalyse, die Lesbarkeit eines Textes, die Wahrscheinlichkeit von Spam usw. Weitere Beispiele sind die Produktivität eines Autors oder die semantische Nähe zwischen einem textuellen Begriff (z.B. einem Suchterm) und einer Zeitschrift. Zwar sind viele Metadaten, die für die Suche genutzt werden, grundsätzlich als Text zu verstehen und können auch so durchsucht werden. Dennoch wird in diesem Artikel speziell auf ihre nicht-textuellen Eigenschaften, etwas solche, die aus informetrischen Analysen stammen, eingegangen.

3.1 Sortierung vs. Ranking

In diesem Abschnitt wird als kurzer Einschub die begriffliche Unterscheidung zwischen Sortierung und Ranking erläutert. Technisch betrachtet sind Sortierung und Ranking einer Dokumentmenge zunächst einmal das Gleiche: Die unsortierte Menge wird in eine geordnete Liste überführt.

Ganz allgemein bedeutet dies, dass eine beliebige Dokumentmenge D, die aus einer Menge von Dokumenten $d \in D$ besteht, in eine linear geordnete Liste L mittels einer Sortierfunktion f überführt wird, sodass gilt:

$$f: D \rightarrow L.$$

Das Ordnungskriterium ist dabei zunächst nicht von Bedeutung und frei wählbar. Dies kann eine lexikographische, alphabetische Ordnung nach Autorennamen oder eine numerische Ordnung nach Dokumenten-ID sein. Es kann aber auch nach einem abstrakten Konzept der Relevanz sortiert werden. Die Rankingfunktion r weist dabei jedem Dokument d einen Wert τ, den sogenannten Relevanzwert (*relevance score*), zu. Die Rankingfunktion r kann wie folgt definiert werden:

$$r(d) \rightarrow \tau.$$

Die Relevanzwerte können geordnet werden, sodass gilt:

$$\tau_1 < \tau_2 \Rightarrow \tau_1 \text{ ist weniger relevant als } \tau_2.$$

Das endgültige Ranking ist dann die Ordnung der Menge D entlang der Relevanzwerte:

$$(\tau_1, \dots, \tau_2) | \forall \tau: \tau_{n-1} < \tau_n.$$

Es macht aus rein technischer Sicht keinen Unterschied, ob eine numerische Ordnung auf Grundlage von Dokumenten-IDs oder berechneten Relevanzwerten hergestellt wird. Die Begriffe werden auch in vielen Systemen und Lehrbüchern synonym verwendet. Dem Benutzer wird angeboten, seine Ergebnisliste nach unterschiedlichen Kriterien zu sortieren, wobei dies sowohl einfache Kriterien wie die alphabetische oder numerische Sortierung, aber auch die Sortierung nach Relevanz sind. Relevanz ist der eigentliche Schlüsselbegriff, der ein simples Sortieren von einem Relevanzranking (*relevance ranking*) unterscheidet.

3.2 Der Relevanzbegriff

Saracevic (2007) beschreibt in seiner Aufsatzreihe die Relevanz als den Kernbegriff der Informationswissenschaft und des Information Retrievals. Der Relevanzbegriff ist ein fundamentales Konzept des Information Retrievals, da die Hauptaufgabe eines IR-Systems darin besteht, relevante Informationen zu liefern (Song u.a. 2011). Die Definition der Relevanz ist vielschichtig und wurde von Borlund (2003) in Form einer ausführlichen Literaturstudie zusammengefasst. Sie führt ein allgemeines Modell der Relevanz in die Diskussion ein, das die verschiedenen Aspekte der Relevanzbewertung zusammenbringt, wie unterschiedliche Klassen, Typen, Grade, Kriterien und Ebenen der Relevanz. Borlund bezeichnet dies als die Multidimensionalität der Relevanz („*multidimensionality of relevance*“). Die allgemeinste Unterscheidungsform der Relevanz ist dabei die Aufteilung in objektive bzw. systembedingte, und subjektive bzw. menschliche Relevanzkriterien.

Borlund (2003, 913) betrachtet vor allem die drei zentralen Schlussfolgerungen über den Relevanzbegriff als solchen:

- Relevanz als ein *multidimensionales, kognitives Konzept*, das von der Wahrnehmung des Benutzers und dessen persönlichem Informationsbedürfnis abhängt;
- Relevanz als ein *dynamisches Konzept*, das von der Bewertung der Beziehung zwischen Information und Informationsbedürfnis zu einem bestimmten Zeitpunkt abhängt;
- Relevanz als ein *komplexes Konzept*, das trotz allem systematisch und messbar erfassbar ist, wenn es dabei aus der jeweiligen Nutzerperspektive betrachtet wird.

Alle drei Aspekte des Relevanzbegriffes eint die Abhängigkeit von einer Nutzerperspektive bzw. einem subjektiven Informationsbedürfnis. Da diese Sichtweise schlecht in Algorithmen und tatsächliche IR-Systeme umsetzbar ist, wird allgemein eine Aufteilung angewendet. Er fasst vier unterschiedliche Arten der Relevanz zusammen:

- *Situative Relevanz*: Hiermit ist die tatsächliche Nützlichkeit einer Informationseinheit in einer konkreten Situation gemeint.

- *Subjektive Relevanz* (bzw. Pertinenz): Die Nützlichkeit einer Informationseinheit für eine Person mit einem bestimmten Informationsbedürfnis.
- *Objektive Relevanz*: Im Unterschied zur subjektiven Relevanz steht hier nicht die subjektive Einschätzung einer Person im Vordergrund, sondern mehrere Einschätzungen unterschiedlicher Personen. Die subjektiven Einschätzungen des Einzelnen sollen durch die Mehrfachbewertungen verschiedener sogenannter Assessoren, z.B. im Rahmen einer IR-Evaluation, abgeschwächt werden.
- *Systemrelevanz*: Die algorithmische Berechnung der Relevanz durch ein IR-System. Hierbei handelt es sich nur um einen Schätzwert des Systems, der auch mit dem Begriff RSV (*Retrieval Status Value*) abgekürzt wird.

Die Aufgabe eines IR-Systems ist es nun, eine auf den berechneten RSV-Werten gerankte Liste für den Suchenden, gemäß seinen Ansprüchen an die situative und subjektive Relevanz, zusammenzustellen.

White (2007, 538-42) betrachtet den Begriff aus einer etwas anderen Perspektive, da er sich der Relevanz über die Relevanztheorie nach Sperber und Wilson annähert, die aus der Position der linguistischen Pragmatik heraus die Relevanz beurteilen. Vor diesem Hintergrund ist die Relevanz für White keine eindeutige Entscheidung, sondern das Verhältnis von kognitivem Effekt auf den Suchenden (*cognitive effect*) zum Verarbeitungsaufwand bzw. zur Verarbeitungsleichtigkeit (*ease of processing*). Diese menschliche Sicht auf die Relevanz (subjektive Sicht) wird nun von ihm auf die Systemsicht abgebildet, die mit Mechanismen wie TF*IDF arbeitet. Er bildet dabei den kognitiven Effekt (d.h. Güte und Mehrwert für den Informationssuchenden) auf TF und die Verarbeitungsleichtigkeit auf IDF ab.

Tab. 3: Abbildung von Systemsicht und menschlicher Sicht auf die Faktoren TF und IDF und den Relevanzeffekte nach White (2007)

	Systemsicht (gemessen)	Menschliche Sicht (nicht gemessen)
TF	Vorhergesagter kognitiver Effekt	Tatsächlicher kognitiver Effekt
IDF	Vorhergesagte Leichtigkeit der Verarbeitung	Tatsächliche Leichtigkeit der Verarbeitung
TF*IDF	Vorhergesagte Relevanz	Tatsächliche Relevanz

Neben der Verknüpfung von kognitivem Effekt und Verarbeitungsleichtigkeit als Messgrößen für die Relevanzbeurteilung stellt White mit den sogenannten Pennant-Diagrammen eine Visualisierung seiner These vor. Es handelt sich dabei um doppelt logarithmierte Streudiagramme (*scatter plots*) der beiden Komponenten TF und IDF. Diese Art der Darstellung hat den Vorteil, dass die unterschiedlichen Komponenten, die zur Bildung des TF*IDF-Wertes führen, besser überblickt werden können. Durch die besondere Darstellung der Pennants lassen sich nun unterschiedliche Zonen A, B und C markieren. Die Zeitschriften in Zone A sind Fachzeitschriften, die als einschlägige Quellen für

dieses Forschungsgebiet angesehen werden können, wohingegen die Zeitschriften der Zone C sich nur peripher mit dem Thema „Lubrication" auseinandersetzen. Dies ist analog zu den Beobachtungen von Bradford zu sehen, die später genauer besprochen werden.

Festzuhalten aus der Pennantdiskussion ist, dass beim Einsatz von TF*IDF die sogenannte „*topical relevance*" überwiegt (White 2007, 547), was in vielen Anwendungsfällen auch genau der intendierte Effekt eines Informationssuchenden ist. Allerdings wird in der Diskussion klar, dass es auch andere Relevanzkriterien gibt, die zwar implizit auch im Pennant zu sehen sind. In einer typischen Ergebnisliste eines IR-Systems sind diese Feinheiten allerdings unter den ersten 10, 20 oder 100 Treffern nicht zu erkennen. Durch die Einteilung in Zonen werden die unterschiedlichen Arten der Relevanz eher sichtbar und erlauben eine erweiterte Sicht auf das Themenfeld. White schließt daraus, dass der Einsatz von TF*IDF die leicht erfassbare Relevanz bevorzugt (z.B. durch Gleichheit von Anfragetermen und Termen im Titel eines Dokuments).

Die Unterscheidung der zwei Begriffe Relevanz und Qualität ist zentrales Anliegen der Arbeiten von Mandl (2006, 93-5). Wie er ausführt, sei für die Bewertung der Relevanz immer ein konkretes Informationsbedürfnis bzw. ein „Informationsproblem" notwendig, zu dem eine konkrete Informationseinheit relevant oder nicht relevant sein könne. Bei Qualität handelt es sich hingegen um ein Attribut, das unabhängig von einem konkreten Informationsproblem zugewiesen werden kann. Allerdings ist eine Qualitätsbewertung nicht objektiv, eine Systemsicht wie beim Relevanzbegriff faktisch nicht umsetzbar. Subjektive Einschätzungen seien aber jederzeit auch ohne konkreten Informationsbedarf möglich. Trotzdem überlappen sich die Begriffe häufig in ihrer Bedeutung und Interpretation. Von den zuvor beschriebenen drei Ebenen der Relevanz nach Borlund ist das Qualitätsurteil dabei nicht zwangsläufig abhängig. Gerade der fehlende Bezug zu einem Informationsbedürfnis macht den Unterschied klar: So mag ein Dokument, das bereits bekannt ist, nicht mehr relevant sein, da es das Informationsbedürfnis nicht weiter befriedigen kann. Dennoch behält das Dokument trotz des verlorenen Neuigkeitswerts nach wie vor seine Qualität.

Gleichzeitig scheint der Relevanzbegriff ein grundlegendes Konzept menschlicher Kognition zu sein:

> People understood and understand relevance similarly over time, space, cultures, and domains. ‚Nobody has to explain to users of IR systems what relevance is, even if they struggle (sometimes in vain) to find relevant stuff. People understand relevance intuitively' (Saracevic 2007, 1918).

Jeder Mensch ist folglich intuitiv in der Lage, ein Urteil über die Relevanz einer Informationseinheit, wie z.B. einem Dokument, zu fällen, wobei dabei immer sein persönliches Informationsbedürfnis im Vordergrund steht. Für ein IR-System besteht nun die Herausforderung darin, diese sehr subjektive Sicht auf ein Informationsproblem aus der Systemsicht heraus zu beantworten.

Die vorangegangene Diskussion zeigt, dass die Relevanzbewertung und damit die Schlussfolgerungen für das Relevanzranking mehrschichtig und vage sind. Während der klassische TF*IDF-Ansatz die themenbezogene Relevanz einfach und deutlich darstellen kann, vernachlässigt er latent relevante oder versteckte Konzepte, die erst durch alternative Verfahren zum Vorschein kommen. Für diese Arbeit ist es daher eine besondere Motivation, nach alternativen Relevanzkonzepten und deren Anwendungen bzw. deren Berechenbarkeit für das Information Retrieval zu suchen. Die nächsten Abschnitte widmen sich daher den unterschiedlichen Systementwürfen zur automatischen Relevanzbewertung durch IR-Systeme und dem dadurch angebotenen Relevanzranking.

3.3 Textuelles Ranking

Im folgenden Abschnitt werden die vier einflussreichsten Modelle, die den aktuellen Stand der Technik darstellen, skizziert und eingeführt. Die hier vorgestellten Verfahren sind in modernen Retrievalplattformen wie Apache Lucene/Solr[12], der Terrier IR Platform[13] oder dem Lemur/Indri Toolkit[14] verfügbar.

Die Modelle sind einer sogenannten Bag-of-Words-Annahme unterworfen, was bedeutet, dass alle Anfragen wie auch Dokumente nur als unzusammenhängende Menge an Wörtern verstanden werden. In dieser Menge von Wörtern sind die Wortreihenfolgen verloren gegangen, was zwar für die Verarbeitung der Anfragen und der Dokumentbasis enorme Vorteile, im Detail aber semantische Probleme mit sich bringt. Hinzu kommen typische Probleme aus dem Natural Language Processing (NLP), u.a. Polyseme, Synonyme oder Wortkomposita. Generell ist die Verwendung der natürlichen Sprache mit Vagheit verbunden, sodass dieses Problem auch als das Vagheitsproblem bekannt ist.

3.3.1 Einfaches und erweitertes boolesches Modell

Das boolesche Modell ist das älteste und meist-unterstützte Retrievalmodell, das auf der Verwendung der booleschen Algebra aufbaut. Eine Anfrage wird in Form eines booleschen Ausdrucks mit Hilfe der Operatoren UND, ODER sowie NICHT ausgedrückt. Durch eine Klammerung der Anfrageterme und der Operatoren lassen sich Anfragen dabei sehr lang und präzise formulieren. So kann mittels der booleschen Algebra eine sehr große Ausdrucksstärke erreicht werden. Dies ist einer der Hauptgründe dafür, dass die boolesche Suche besonders bei geschulten Nutzern und Experten beliebt ist. Sie erlaubt es, sowohl eine präzise Anfrage zu stellen, als auch das Ergebnis transparent und kontrollierbar zu gestalten (Manning, Raghavan und Schütze 2008, 3-5). Charakteris-

[12] <http://lucene.apache.org> (Zugegriffen: 12. Oktober 2013).
[13] <http://terrier.org> (Zugegriffen: 12. Oktober 2013).
[14] <http://www.lemurproject.org> (Zugegriffen: 12. Oktober 2013).

tisch für das boolesche Retrieval ist eine häufige Neu- und Weiterformulierung der Ursprungsanfrage, z.B. um Synonyme zur Suche hinzuzufügen. Boolesches Retrieval unterstützt nur exakte Anfragen, die auf dem Exact-Match-Verfahren basieren. Studien haben allerdings auch immer wieder gezeigt, dass die Mächtigkeit der booleschen Logik nur von wenigen Nutzern korrekt verstanden und angewendet wird, wie z.B. in einer Studie von Zhang (2008) zum Suchverhalten von Universitätsstudenten gezeigt. Ein Gros der Nutzer ist mit der korrekten Anfrageformulierung überfordert.

Die einfache technische Umsetzung mit Hilfe eines invertierten Index, die Mächtigkeit der Anfrageformulierung sowie die immense Verbreitung in unzähligen, digitalen Bibliotheken sind wohl die größten Vorteile des Verfahrens. Allerdings sind auch die größten Nachteile des Systems offensichtlich: Durch das Exact-Match-Verfahren werden viele relevante Dokumente nicht gefunden; die korrekte Anfrageformulierung ist nicht intuitiv und fehleranfällig. Darüber hinaus beinhaltet das klassische boolesche Retrieval kein Ranking. Lange Ergebnislisten können somit nur nach einfachen Dokumenteneigenschaften sortiert werden. Da dies bei großen Dokumentmengen zu Problemen führt, wurde das boolesche Retrieval erweitert.

3.3.2 Erweitertes boolesches Ranking

Eine Erweiterung des Konzeptes ist als Extended-Boolean-Retrieval bekannt. Hierbei werden unterschiedliche Gewichte in das Modell, wie z.B. Termhäufigkeiten, eingebracht. Weiterhin können durch die Erweiterungen unterschiedliche, strukturelle Informationen der Dokumente besser verarbeitet werden. So kann unterschieden werden, ob Terme z.B. nur im Titel eines Dokuments vorkommen oder z.B. auch im Abstract. Gerade in Verbindung mit den reichen Beständen an Metadaten entsteht so eine mächtige Anfragesprache, die bis heute in vielen digitalen Bibliotheken Bestand hat. Eine der einfachsten Varianten, ein erweitertes boolesches Retrieval zu implementieren, ist, die Anfrageterme nach der bereits erwähnten Termhäufigkeit in den Dokumenten zu gewichten. Die Termhäufigkeit wird auch Termfrequenz (*term frequency*) genannt und für einen Term t in einem Dokument d gilt: $\mathrm{tf}(t, d)$. Je häufiger ein Anfrageterm in einem Dokument vorkommt, desto relevanter ist er und verschafft dem Dokument eine höhere Position in der Ergebnisliste.

Die Termfrequenz ist zwar ein guter Indikator für die Wichtigkeit von Termen in einem Dokument, doch steigt die Relevanz eines Textes nicht proportional mit der Häufigkeit von Termnennungen an. Um dieses Ungewicht auszugleichen, wird die inverse Dokumenthäufigkeit (*inverse document frequency*) hinzugefügt (Spärck Jones 1972):

$$\mathrm{idf}(t) = \log \frac{N}{1 + \mathrm{df}(t)}$$

N enspricht dabei der Anzahl der gesamten Dokumentmenge und $\mathrm{df}(t)$ ist die Anzahl der Dokumente, die den Term t enthalten. Um eine Division durch 0 zu verhindern, wird dem Divisor üblicherweise eine 1 hinzuaddiert. Beide Werte können zum sogenannten TD*IDF-Gewicht (*term frequency inverse document frequency*) zusammengefasst werden:

$$\mathrm{TF*IDF} = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t) = \mathrm{tf}(t,d) \cdot \log \frac{N}{1 + \mathrm{df}(t)}.$$

TF*IDF weist einem Term t in einem Dokument d einen Wert zu, der vom unterschiedlichen Auftreten von t und d abhängt. Dieser Wert ist ein statistisches Maß, das zeigt, wie wichtig ein Dokumententerm in einer Dokumentmenge ist. Der Wert ist am höchsten, wenn t häufig in einer geringen Anzahl an Dokumenten auftritt. Man spricht dabei von der diskriminierenden Kraft der Terme für die Dokumentmenge. Der Wert ist niedrig, wenn t entweder selten in einem Dokument vorkommt oder wenn t in einer großen Anzahl von Dokumenten enthalten ist. Je häufiger ein Term in vielen Dokumenten enthalten ist, desto weniger ausdrucksstark ist er. TF*IDF ist daher am niedrigsten, wenn der Term t in allen Dokumenten der Dokumentmenge enthalten ist. In diesem Fall hat er keine besondere diskriminierende Kraft mehr. Dies tritt vor allem bei Wörtern wie „und“, „ist“ oder Artikeln auf. Diese Wörter können z.B. über eine Stoppwortliste aus dem Verfahren hinausgefiltert werden.

Werden nun die TF*IDF-Werte für alle Terme t in einer Anfrage q aufsummiert, kann für jedes Dokument d der Dokumentmenge ein eigener Relevanzwert berechnet werden. Eine Sortierung der Dokumentmenge anhand dieses Wertes (auch Score genannt) ist eine direkte Anwendung eines Relevanzrankings:

$$\mathrm{score}_{\mathrm{TF*IDF}}(q,d) = \sum_{t \in q} \mathrm{tf}(t,d) \cdot \mathrm{idf}(t).$$

Viele Varianten dieses ursprünglichen Modells sind bekannt, so z.B. Abwandlungen, die eine Normalisierung der Dokumentlänge oder eine Skalierung der tf-Werte berücksichtigen. Der eigentliche Ansatz, dass abhängig von der Anfrage jedem Dokument ein Score zugewiesen wird, wird nicht verändert. Ein Beispiel hierfür wäre die Möglichkeit, mittels logarithmischer Skalierung des tf-Wertes Einfluss auf die Gewichtung zu nehmen. Die grundsätzliche Annahme, dass ein n-faches Auftreten des Terms t einer n-fachen Steigerung der Systemrelevanz gleichkommt, wird durch eine Logarithmierung der Termfrequenz tf abgeschwächt.

Das erweiterte boolesche Retrieval repräsentiert in diesem Fall ein Dokument als einen Vektor und arbeitet mit Termgewichtungen. Dieser Ansatz bildet den Schnittpunkt zum Vektorraummodell, welches im folgenden Abschnitt vorgestellt wird.

3.3.3 Vektorraummodell

Eines der Standardverfahren zur Implementierung eines Relevanzrankings in IR-Systemen ist das Vektorraummodell, das gleich mehrere Schwächen des booleschen Modells auszugleichen versucht: zum einen wird ein Ranking eingeführt, zum anderen kann durch die geometrische Interpretation der Dokumente die Strenge der booleschen Logik entschärft werden; auch Dokumente, in denen nicht alle Anfrageterme vorkommen, können in der gerankten Ergebnisliste vorkommen und eine hohe Platzierung erreichen, da sie trotzdem relevant sein können (Robertson 1977).

Im Gegensatz zum einfachen booleschen Modell werden dabei sowohl die Dokumente selbst, als auch die Anfragen als Vektoren verstanden. Der Vektor $\vec{V}(q)$ ist der Anfragevektor und der Vektor $\vec{V}(d)$ ist der Dokumentvektor. Innerhalb dieser Vektoren entspricht jeder Term einer Dimension. Die unterschiedlichen Terme eines Dokuments oder einer Anfrage spannen so einen mehrdimensionalen Raum auf. Da es sich beim Vektorraummodell um einen klassischen Bag-of-Words-Ansatz handelt, wird die Reihenfolge der Terme beim Aufspannen der Vektoren ignoriert. Gesucht werden nun die Dokumentenvektoren, die dem Anfragevektor ähnlich bzw. nahestehend sind. Dies kann z.B. über den Kosinus-Abstand bestimmt werden, der sich aus dem Skalarprodukt geteilt durch das Produkt der Länge der beiden Vektoren ergibt:

$$\text{score}_{\text{cosine}}(q,d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{\left|\vec{V}(q)\right|\left|\vec{V}(d)\right|}$$

In praktischen Umsetzungen werden die beiden Vektoren mit gewichteten Werten gefüllt, wobei $\text{score} \in \mathbb{R}^{|V|}$ gilt. Das TF*IDF-Gewichtungsmaß wird hierbei am häufigsten angewendet.

3.3.4 Probabilistisches Modell

Neben der TF*IDF-Rankingfunktion ist Okapi BM25 oder die aktuellere Implementation BM25F (für strukturierte Dokumente) bei textuell-orientierten Retrievalkampagnen wie TREC sehr erfolgreich. Das System ist nach dem Okapi-System der City University London benannt und stellt die 25. Iteration einer Reihe von Best-Match-Systemen dar (Robertson u.a. 1995). Obwohl BM25 auf dem sogenannten probabilistischen Modell basiert und so folglich eine andere theoretische Grundlage als das Vektorraummodell besitzt, kann es doch auch wie TF*IDF zu den klassischen Bag-of-Words-Ansätzen gezählt werden. Auch die Berechnung des BM25-Scores ist abhängig von den Anfragetermen und rankt die Ergebnisse einer Treffermenge entsprechend der ermittelten Relevanz zwischen Anfragetermen und der Termverteilung in der Dokumentenbasis. Für eine Anfrage q, welche die Terme t enthält, wird der BM25-Score wie folgt berechnet:

$$\text{score}_{\text{BM25}}(q,d) = \sum_{t \in q} \text{idf}(t) \cdot \frac{(k+1) \cdot \text{tf}(t,d)}{\text{tf}(t,d) + k\left((1-b) + b\frac{dl}{avdl}\right)},$$

wobei idf die inverse Dokumentfrequenz, tf die Termfrequenz sowie dl und avdl die aktuelle Dokumentlänge bzw. die durchschnittliche Dokumentlänge in der gesamten Dokumentmenge sind. Die beiden Parameter $0 \leq k$ und $0 \leq b \leq 1$ werden zur Normalisierung genutzt und meist auf feste Werte gesetzt (Robertson und Zaragoza 2010, 360).

Das probabilistische Modell beruht auf bedingten Wahrscheinlichkeiten und macht in der zuvor skizzierten Umsetzung Annahmen über die Unabhängigkeit der verwendeten Terme sowie deren Einfluss auf die eigentliche Relevanz eines Dokuments. Die grundsätzliche Idee hinter BM25 und dem probabilistischen Modell ist dabei aber auch die Vagheit der Sprache, die bei der Anfrageformulierung besteht und in das Modell eingearbeitet wurde. So wird im probabilistischen Modell nur die Wahrscheinlichkeit für die spätere Relevanz berechnet. Obwohl Relevance-Feedback (die Einbeziehung von Relevanzbewertungen durch den Benutzer) sowohl für das Vektorraummodell, als auch für das probabilistische Modell als Konzept existiert, fügt es sich doch speziell im letzten besonders nahtlos in die dahinterstehende Theorie ein. Dies ist auch einer der Vorteile, die von Fuhr (2004, 211) skizziert werden: Das Modell kann sowohl über die Retrievalmaße, als auch entscheidungstheoretisch über die Kosten begründet werden. Trotzdem ist das Verfahren jenseits von Evaluationskampagnen kaum in kommerziellen Systemen oder Websuchmaschinen vorzufinden.

Sonderformen der probabilistischen Modelle sind die sogenannten statistischen Sprachmodelle (*Language-Models*). Sie stellen formale Beschreibungen einer Sprache dar, die bei der Erzeugung von Texten den Regeln dieser Sprache entsprechen. Die Idee, Sprachmodelle für das IR einzusetzen, basiert darauf, dass sich zu einem existierenden Text über die darin enthaltenen Wörter ein passendes Modell ableiten lässt. Für den Einsatz in IR-Systemen kann man somit zu jedem Dokument ein entsprechendes Modell generieren und zu einer Anfrage prüfen, welches Dokumentmodell die Anfrage am ehesten erzeugen kann.

3.3.5 Vergleich der Modelle

Mit dem Vektorraummodell, dem probabilistischen Modell und den statistischen Sprachmodellen existieren leistungsfähige Alternativen zum einfachen bzw. erweiterten booleschen Retrieval. Diese Modelle haben sich in den Evaluationskampagnen wie TREC oder CLEF beweisen können und stellen in der Forschung den aktuellen Stand der Technik dar.

Betrachtet man allerdings den derzeitigen Stand der Technik bezüglich der Retrieval- und Rankingmodelle in digitalen Bibliotheken, so herrscht nach wie

vor das erweiterte boolesche Retrieval vor. So arbeiten auch große Systeme wie Scopus, Web of Science oder auch RePeC mit einem einfachen, term-basierten Relevanceranking oder, wie z.B. arXiv, vollständig ohne ein Ranking. Dies liegt zum Teil daran, dass die alternativen Modelle ihre jeweiligen Stärken nicht voll ausspielen können. Die Retrievalleistungen der unterschiedlichen Ansätze sind je nach Anwendungsgebiet vergleichbar, wenn eine jeweilige domänenspezifische Anpassung und ein Tuning der Parameter sowie entsprechende Vor- und Nachverarbeitungsschritte eingeplant wurden. Hier werden teilweise mit viel händischer Arbeit die jeweiligen Modelle um weitere Schritte und Faktoren erweitert, dem sogenannten *Feature Engineering* (Manning, Raghavan und Schütze 2008, 311), wobei der Begriff im Bereich der Dokumentklassifizierung geprägt wurde und nicht so häufig im IR angewendet wird. Die eigentliche Retrievalleistung der zugrundeliegenden Rankingmodelle tritt dabei zusehends in den Hintergrund. Es kann folglich keines der Modelle per se als das Beste angesehen werden (Metzler 2011, 22).

Dies zeigte sich auch in Studien von Zhu et al. (2009, 707), in denen die besten MAP-Werte eines TF*IDF-, BM25- bzw. LM-basierten Systems nur minimale Unterschiede zeigten (TF*IDF = 0,2233, BM25 = 0,2220 und LM = 0,2244). In dieser Arbeit wurden auf Grundlage der TREC2006 Expert Search Test Collection die Rankingmodelle mit internen Dokumentstrukturen, PageRank und einer Anfrageerweiterung angereichert, jedoch zeigten sich sowohl bei der Anwendung der unveränderten Rankingmodelle, als auch den besten Kombinationen keine signifikanten Unterschiede (TF*IDF+QE+PR = 0,4081, BM25+QE+PR = 0,4067 und LM+QE+PR = 0,4087).

Ähnliche Ergebnisse zeigten sich in einer Metastudie von Kürsten und Eibl (2011), in der fünf unterschiedliche Stemmer, dreizehn IR-Modelle und drei Pseudo-Relevance-Feedback-Mechanismen in allen möglichen Kombinationen miteinander verglichen wurden. Es zeigte sich weiterhin, dass gerade im Zusammenspiel der verschiedenen Systemkomponenten die Wahl des IR-Modells nur ein Faktor unter vielen ist. Konkret konnte eine TF*IDF-Implementation das zweitbeste Ergebnis hinter einer Kombination mit einem probabilistischen und parameterfreien DLH-Modell (eine Variante des Divergence from Randomness-Modells) erlangen. Die arithmetisch gemittelten MAP-Werte unterschieden sich mit 0,3578 an erster (DLH) und 0,3518 an zweiter Stelle (TF*IDF) nur knapp.

Dies deckt sich mit der Einschätzung von Fang Tao und Zhai:

> The state-of-the-art retrieval functions, when optimized, usually have similar MAP values even though their function forms are different and their retrieval results for the same query also tend to differ. This suggests that all the functions may have their own (potentially different) weaknesses and strengths (Fang, Tao und Zhai 2011, 3).

3.4 Nicht-textuelles Ranking

Alle bisher vorgestellten Retrieval- und Rankingmodelle basieren auf dem zuvor beschriebenen Bag-of-Words-Ansatz: Anfrage- und Dokumentterme werden dabei verglichen und jeweils der Exact- oder Best-Match dem Benutzer in Form einer gerankten Ergebnisliste angeboten. Die Bewertung der jeweiligen Relevanz erfolgt ausschließlich auf Grundlage von Termhäufigkeiten oder anderen ermittelten sprachlichen Eigenschaften. In dieser HSR Focus werden diese als textuelle Dokumentfeatures bezeichnet. Dokumente in digitalen Bibliotheken bestehen nicht nur aus textuellen Features und es lassen sich aus der Fülle an Metadaten weitere Features ermitteln, die ein Dokument und seine Relevanz bzgl. einer Nutzeranfrage beschreiben. Neben den textuellen Eigenschaften, die sich direkt aus den Dokumenten auslesen lassen, können zusätzliche Daten aus und über die Dokumente, ihre Autoren oder sonstige Metadaten verwendet werden, um ein Ranking umzusetzen. Für Ferber (2003, 302-6) sind hierbei vor allem die Angaben von Interesse, die von „(möglichst fachkundigen) Menschen gemacht werden, also manuelle Indexierungen, Beschreibungen und Beurteilungen oder Verweise [...]".

Im folgenden Abschnitt sollen genau diese Eigenschaften vorgestellt werden. Im Einzelnen sind dies u.a. Verfahren aus der Websuche wie der PageRank oder auf Zitationsanalyse beruhende, wissenschaftliche Rankingverfahren wie der h-Index. Eine Mischform aus beiden Ansätzen stellt der Google-Scholar-Rank dar, der, zusammen mit anderen aus der Welt der Netzwerkanalyse entlehnten Ansätzen, wie z.B. der Autorenzentralität oder der Popularitätsbestimmung in sozialen Netzwerken im Web 2.0, vorgestellt wird.

3.4.1 PageRank und HITS

In der Frühzeit des Web waren traditionelle term-basierte Retrievalmethoden auf Basis der zuvor beschriebenen Verfahren wie dem booleschen oder dem Vektorraummodell verbreitet. Angereichert wurden diese zwischenzeitlich durch Ansätze wie dem Sponsored Search, bei dem primär die Höhe der Zahlung des Webseitenbetreibers in das Ranking einging und Suchmaschinenbetreibern wie Overtune zu kommerziellem Erfolg verhalf. Mit dem PageRank wurde nun erstmals ein Verfahren zur Suche im Web entwickelt, das primär auf ein nicht-textuelles Rankingverfahren setzte (Page u.a. 1999).

Ferber (2003, 302) beschreibt den PageRank als eine rekursive Funktion:

$$\mathrm{PR}(i) = (1-d) + d\left(\frac{\mathrm{PR}(1)}{\mathrm{out}(1)} + \cdots + \frac{\mathrm{PR}(m)}{\mathrm{out}(m)}\right).$$

Der PageRank einer Seite $\mathrm{PR}(i)$ wird aus den PageRank-Werten der Webseiten berechnet, die auf die aktuelle Webseite zeigen. Mittels des Wertes out wird die Zahl der ausgehenden Links einer Seite angegeben. Durch die Division mit

diesem Wert erfolgt ein Ausgleich zwischen Webseiten mit vielen und solchen mit wenigen ausgehenden Links (*Outlinks*). Der Parameter $0 \leq d \geq 1$ kann als Dämpfungsfaktor verwendet werden. Durch die Addition aller PageRanks der Seiten, die auf die aktuelle Seite i verweisen, zuzüglich dem Dämpfungsfaktor und dem Wert 1, kann der PageRank der aktuellen Seite bestimmt werden.

Die eigentliche Idee des PageRanks basiert auf der sogenannten Random Surfer-Annahme (Page u.a. 1999, 3), womit das Verfahren inhaltlich motiviert werden kann. Bei der praktischen Umsetzung werden dazu die Anzahl der Backlinks – die eingehenden Links auf eine Webseite – gezählt und gleichzeitig auf die unterschiedliche Wertigkeit von Links Bezug genommen. Eine Webseite wird dabei als wichtig oder einflussreich bezeichnet, wenn andere wichtige Seiten auf sie zeigen. Es macht beim PageRank-Verfahren einen großen Unterschied, ob die eingehenden Links von wichtigen oder von weniger wichtigen Webseiten stammen. Es zählt nicht nur die schiere Menge an Verlinkungen, sondern ebenso die Wichtigkeit der verlinkenden Webseite.

Der so ermittelte Wert pro Seite ist dabei vollkommen unabhängig von der jeweiligen Suchanfrage. Es existiert so pro Webseite jeweils ein PageRank-Wert, der global gilt und statisch ist. Zwar kommen durch die Dynamik des Webs ständig neue Webseiten in die Graphenstruktur hinzu, doch wird davon ausgegangen, dass die Links zwischen den Seiten sich nicht innerhalb kurzer Zeit ändern, also stabil sind. Es muss folglich nicht zwangsläufig für jede neue Webseite eine Neuberechnung des PageRanks angestoßen werden. Durch die Unabhängigkeit von der Suchanfrage ist auch die konkrete Umsetzung innerhalb der Websuchmaschine so angelegt, dass zunächst mit Hilfe einer erweiterten booleschen Anfrage eine Dokumentmenge mit potentiell relevanten Dokumenten ermittelt wird, die anschließend mittels des globalen PageRanks gerankt wird. Page und Brin bezeichnen ihr Verfahren dabei so, dass mit der term-basierten Anfrage für eine hohe Precision und mittels des PageRanks für eine besonders hohe Qualität am oberen Ende der Ergebnisliste gesorgt wird. Die Grundannahme eines globalen Rankings von Webseiten, unabhängig von der jeweiligen Thematik der Suche und dem dadurch unterschiedlichen Informationsbedürfnis eines jeweiligen Nutzers, ist allerdings nicht unumstritten.

Kleinberg (1999) stellte ein alternatives Verfahren vor, das er den HITS-Algorithmus nannte (*Hypertext-Induced Topic Selection*). Wie der Name bereits sagt, sollte hier die Schwäche des PageRanks (eines undifferenzierten globalen Rankings über alle Themengebiete hinweg) umgangen werden. Hierzu werden zwei unterschiedliche Arten von Webseiten eingeführt, die sogenannten *Hubs* (Mittelpunkte) und *Authorities* (Autoritäten). Authorities sind Webseiten, die gute, einschlägige Inhalte besitzen. Hubs sind andere Webseiten, die auf diese Inhalte verlinken. Hubs und Authorities stehen in einer wechselseitigen Beziehung, da sie sich gegenseitig in ihrer Wichtigkeit (wie im Sinne des PageRanks) bestärken.

Bei HITS werden jeder Webseite zwei Werte zugewiesen, der Hub- und der Authorities-Wert, jeweils berechnet pro Anfrage. Dies steht im klaren Gegensatz zu PageRank, welcher anfrageunabhängig berechnet wird. Gleichzeitig werden die beiden Werte nicht auf der gesamten Dokumenten/Webseitenmenge berechnet, sondern nur auf einer kleinen Untermenge an zuvor als relevant befundenen Webseiten. Hierzu wird ein einfaches textuell-basiertes Retrievalverfahren verwendet (im Papier von Kleinberg ist dies die Websuchmaschine AltaVista), um die sogenannte Grundmenge (*root set*) zu erzeugen. Mit dieser Grundmenge wird nach einem weiteren Vorverarbeitungsschritt, der die Grundmenge noch einmal erweitert, gerechnet. Durch die verwendeten textuellen Verfahren wird davon ausgegangen, dass die so gebildete Grundmenge größtenteils homogen bzgl. der Thematik der Anfrage ist. Diese Annahme ist vor einer ähnlichen Problematik im Bereich der Query nicht unproblematisch, kann es doch anstelle eines Query-Drift zu einem Topic-Drift kommen. Durch die Erweiterung können nicht-relevante Seiten hohe Authority-Werte erlangen, obwohl sie nicht zur ursprünglichen Suchanfrage passen.

3.4.2 Google Scholar Citation-Rank

Ein sehr prominentes System, das einen komplexeren Rankingalgorithmus unter der Zuhilfenahme nicht-textueller Features implementiert, ist Google Scholar, welches nach eigener Aussage unterschiedliche Features kombiniert „[... in a ...] way researchers do, weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature“ (Google Scholar 2011). Google Scholar setzt auf ein kombiniertes Ranking, das sowohl Termfrequenzen, den Publikationsort, die Autorschaft und die Zitationen des Textes beinhaltet. Wie üblich bei einem kommerziellen Produkt ist die genaue Umsetzung ein Geschäftsgeheimnis, doch zeigen Untersuchungen des Retrievalverhaltens von Beel und Gipp (2009), dass die Zitationsrate und der Abgleich von Anfrage- und Titeltermen den höchsten Einfluss auf das Ranking haben, wobei andere Kriterien wie beispielsweise das Alter der Dokumente keinen großen Einfluss haben. Dies führt dazu, dass die höchst-gerankten Treffer einer Suche in Google Scholar meist Dokumente mit einer hohen Anzahl an Zitationen sind. Anders als Page und Brin, die in ihrem ursprünglichem Artikel zum PageRank-Verfahren gegen die Auszählung von Zitationszahlen (bzw. den reinen Backlinks) argumentierten, setzt Google Scholar sehr stark auf den Einfluss von Zitationen. Google Scholar erlaubt es seinen Benutzern nicht selbst, das Ranking zu beeinflussen oder alternative Sortiermöglichkeiten auszuwählen.

3.4.3 Wissenschaftliches Ranking: Journal Impact Factor und h-Index

Der Journal Impact Factor (JIF) ist ein Maß, das am Institute for Scientific Information (ISI) um Eugene Garfield entwickelt wurde. Der JIF wird jeweils für ein aktuelles Jahr berechnet, indem für die beiden vorgegangenen Jahre die Anzahl der tatsächlich zitierten Artikel aus einer Zeitschrift durch die Anzahl der potentiell zitierbaren Artikel geteilt wird:

$$\mathrm{JIF}(J)_{2012} = \frac{\mathrm{Zitationen}(J)_{2010} + \mathrm{Zitationen}(J)_{2011}}{\mathrm{Zitierbare\ Artikel}(J)_{2010} + \mathrm{Zitierbare\ Artikel}(J)_{2011}}.$$

Sind für Artikel einer Zeitschrift J in den Jahren 2010 und 2011 insgesamt 200 Zitationen registriert worden und die Zeitschrift hat in diesem Zeitraum insgesamt 100 Artikel herausgebracht, dann ist der JIF für das Jahr 2012: 200/100=2,0.

Die Aussagekraft des JIF ist, wie viele andere Indizes auch, umstritten. Speziell bei dem vom ISI herausgegebenen Journal Citation Report (JCR) wird immer wieder die problematische fachliche Abdeckung als Kritikpunkt angeführt. Es ist offensichtlich, dass für die exakte Bestimmung der Zitationen eine möglichst große Abdeckung der verwendeten Fachdatenbank von Nöten ist. Gleichzeitig besteht häufig Unklarheit über die Definition eines „zitierbaren Artikels", da je nach Betrachtungsweise Leitartikel, Editorials, Konferenzbände etc. mit in die Zählung einfließen. Ein weiteres Spannungsfeld stellen Selbstzitationen dar, wie z.B. Glänzel et al. (2008) für die Domäne der Bioinformatik herausstellt. Auch die zeitliche Abdeckung von zwei Jahren wird als problematisch erachtet, da davon ausgegangen werden kann, dass manche Papiere länger benötigen, um von der wissenschaftlichen Community aufgegriffen zu werden. Während der JIF primär für Zeitschriften entwickelt wurde, gibt es zudem Variationen für Webseiten oder Domänennamen im Internet (Web Impact Factor).

Ein anderer Indikator, der sich trotz seines noch jungen Alters als sehr einschlägig erwiesen hat, ist der h-Index von Hirsch (2005). Der h-Index wird zur Bewertung einzelner wissenschaftlicher Autoren, Zeitschriften oder Institutionen herangezogen. Er wird dabei sehr einfach als die Anzahl von h Papieren, die mindestens h-mal zitiert wurden, berechnet. Ein h-Index von 6 bedeutet, dass ein Autor mindestens 6 Papiere geschrieben hat, die mindestens 6-mal zitiert wurden. Redner (2010) zeigt, dass der h-Index in enger Verbindung zur Anzahl der Zitationen steht, sodass $\sqrt{c}$ wie h skaliert, wobei c die Anzahl der Zitationen eines Autors darstellt.

Der h-Index hat sich als sehr robuster Faktor herausgestellt, da Ausreißer wie Publikationen mit einer unverhältnismäßig hohen Zitationsrate keinen Einfluss auf den Wert des h-Index haben. Auch eine große Anzahl an Publikationen, die jedoch selten zitiert wurden, hat keinerlei Einfluss. Gleichzeitig kann aber festgestellt werden, dass die h-Indizes zweier Forscher nur bedingt

miteinander zu vergleichen sind, da z.B. die Dauer der Publikationstätigkeit eines Autors erheblichen Einfluss auf den h-Index hat. Junge Forscher in ihren ersten Publikationsjahren haben keine Chance auf einen hohen Wert. Des Weiteren ist der h-Index, ebenso wie der JIF, fachlich nicht übertragbar, da in den unterschiedlichen wissenschaftlichen Domänen grundverschiedene Publikationstraditionen herrschen.

In Zeiten frei verfügbarer Publikationen im Netz bzw. in digitaler Form über digitale Bibliotheken stellt sich die Frage, ob die reinen Zitationszahlen das Maß aller Dinge sind. Schlögl und Gorraiz (2012) schlagen beispielsweise vor, Zitationszahlen durch Downloadstatistiken zu ersetzen:

> Es stellt sich nun die Frage, ob es nicht geeignetere Datengrundlagen gibt. Anbieten würden sich hier vor allem Download-Daten. Diese fallen in den Bibliotheken selbst an und bringen dadurch das tatsächliche Nutzungsverhalten von wissenschaftlichen Zeitschriften besser zum Ausdruck. Außerdem sind Download-Daten aktuell, während Zitationsdaten erst mit einer mehr oder weniger großen Verspätung anfallen (Schlögl und Gorraiz 2012, 87).

Als Abschluss ihrer Untersuchung stellen die beiden Autoren fest, dass die meistzitierten Zeitschriften auch diejenigen sind, die die höchsten Downloadraten verzeichnen. Einen ähnlichen, aber abgeschwächten Zusammenhang stellen sie auf der Artikelebene und bei den Impactfaktoren fest.

Ist man sich der Nachteile der zuvor vorgestellten Verfahren bewusst, kann man mit Tools wie Publish or Perish[15] von Harzing (2010) auf Grundlage von Daten aus Google-Scholar alle o.g. Werte berechnen.

3.4.4 Zentralität in Autorennetzwerken

Eine andere Art der Netzwerkanalyse ist die Berechnung der sogenannten Autorenzentralität. Die Annahme hinter dieser Berechnung ist, dass sich die Bedeutung eines Dokuments durch den Einfluss seines Autors bestimmen lässt. Dieser Einfluss wird hierbei mit der Zentralität des Autors in einem Netzwerk von Koautoren gleichgesetzt. Die Relevanz eines Dokuments steigt mit der Zentralität seiner Autoren in besagtem Netzwerk. Nutzt man die errechnete Zentralität der Autoren als Sortierkriterium, erhält man ein alternatives Relevanzranking für eine gegebene Dokumentmenge.

Der Hintergrund dieser Überlegung ist, dass in der Wissenschaft die Generierung von neuem Wissen in ein soziales Netzwerk von Wissenschaftlern eingebunden ist. Wissenschaftler arbeiten nicht isoliert von der sozialen Außenwelt, sondern im Austausch mit Kollegen. Die Ergebnisse dieser wissenschaftlichen Arbeit werden auch gemeinsam publiziert. Es bildet sich so ein großes, meist internationales Netzwerk von Koautoren (He 2009).

[15] <http://www.harzing.com/pop.htm> (Zugegriffen: 12. Oktober 2013).

Die Zusammenarbeit in der Wissenschaft ist in ihrer Art und Weise nach verbunden mit dem Konzept der Betweenness-Zentralität, wie sie bei der Analyse von sozialen Netzwerken verwendet wird. Die Betweenness beschreibt den Grad, zu welchem ein Knoten (in diesem Fall ein Wissenschaftler) auf einem kürzesten Pfad zwischen anderen Knoten liegt. Anders ausgedrückt kann dies als der Grad interpretiert werden, zu dem ein Knoten eine vermittelnde Rolle zwischen anderen Knoten einnimmt. Die Autorenschaft mit anderen Autoren sei dabei „a process in which knowledge flows among scientists" (Yin u.a. 2006) und ein Autor mit hoher Betweenness nähme einen entsprechend stark frequentierten Punkt innerhalb dieses Netzwerks ein, der gleichzeitig unterschiedliche Gruppen im Netzwerk miteinander verbindet.

Neben empirischen Studien, welche die Bedeutung des Betweenness-Maßes unterstreichen, wurden die starken Zusammenhänge zwischen Zitationszahlen und Betweenness als Indikator für die weiter Verwendung dieses Maßes für das IR genutzt. Mutschke et al. (2011) machen sich diese Bedeutung zu Nutze, indem sie die Betweenness als ein Rankingkriterium für die wissenschaftliche Informationssuche verwenden. Sie beschreiben das Verfahren als ein mehrschrittiges, das ausgehend von einer initialen Dokumentmenge, die auf Grundlage von Indextermen zusammengestellt wird, ein Reranking auf Betweenness der Autoren anbietet.

3.4.5 Web 2.0-Retrieval und Social-Ranking

Eine andere Situation kann in sogenannten Web 2.0, Social-Bookmarking oder anderen dem Social-Web zuzuordnenden Plattformen beobachtet werden. Hier haben sich in den letzten Jahren auch speziell für den wissenschaftlichen Bereich entsprechende Angebote platzieren können. Beispielhaft für den Bereich der digitalen Bibliotheken seien hier die beiden Plattformen Bibsonomy (Benz u.a. 2010) und CiteULike als von der Wissenschaftsgemeinde getragene Plattformen genannt. Der Inhalt dieser beiden Systeme wird primär von seinen Benutzern erzeugt und gepflegt (*User-generated Content*) und ist daher in Bezug auf die Erschließungsqualität mit traditionellen digitalen Bibliotheken nicht zu vergleichen, da hier ein Großteil der Daten von Laien und daher sehr heterogen erfasst wird. Es kommen keine *einheitlichen Knowledge Organization Systems* (KOS) wie Thesauri oder feste Klassifikationen zum Einsatz, sondern Social Recommendations und bevorzugt Tagging. Diese Art der Knowledge-Organization wird als *Folksonomie* bezeichnet (Peters 2010).

Soziale Netzwerke und Web 2.0-Plattformen bieten eine Fülle an neuartigen, nicht-textuellen Features, die für das Retrieval genutzt werden können. Neben den offensichtlichen sozialen Aspekten wie Empfehlungen oder Popularitätswerten sind dies implizites und explizites Nutzer-Feedback sowie das Nutzungsverhalten. Wie Hotho et al. (2006) zeigen, sind für solche auf Folksonomien basierende Systeme „ordinary ranking schemes such as TF*IDF [...]"

nicht ausreichend. Dies liege u.a. daran, dass die Textteile, die genutzt werden, um Webseiten, Bilder oder wissenschaftliche Dokumente zu beschreiben, nur sehr kurz sind und ein Volltext nicht verfügbar ist. Um dieses strukturelle Problem zu lösen, schlagen die Autoren zwei alternative Rankingverfahren vor, die sie Adapted PageRank und FolkRank nennen. Diese Verfahren sind besonders auf die Eigenarten und Strukturen von Folksonomien ausgerichtet und können auf großen Folksonomie-Datensätzen evaluiert werden.

Der Adapted PageRank teilt mit dem klassischen PageRank die prinzipielle Idee eines Graph-basierten Ansatzes, wobei jeder einflussreiche Knoten den Einfluss anderer einflussreicher Knoten unterstützt, auf die er verlinkt bzw. mit denen er gleichzeitig kookkurriert. Ein Nutzer kann in diesem Netzwerk ein beliebiges Dokument oder andere Ressourcen taggen, wobei das Dokument bzw. die Ressource mehr an Einfluss gewinnt, je einflussreicher der Benutzer ist. Die drei unterschiedlichen Entitäten, die in diesem System verwendet werden, sind Dokumente bzw. Ressourcen, Benutzer und Tags. Die zuvor skizzierte gegenseitige Verstärkung des Einflusses kann symmetrisch auf Tags und Benutzer übertragen werden. Während der Adapted PageRank jeweils ein globales Ranking berechnet, kann der FolkRank ein personalisiertes Ranking und daraus Empfehlungen für den Benutzer generieren. Diese Empfehlungen können Dokumente, Tags oder auch andere Benutzerprofile sein, die von potentiellem Interesse für den jeweiligen Benutzer sein können.

Dieser Ansatz wurde von Zanardi und Capra (2008) aufgegriffen, die die Ähnlichkeit zwischen Benutzerprofilen maßen, indem sie Tag-Gewohnheiten von Nutzern verglichen und dieses Ähnlichkeitsmaß nutzten, um Dokumente zu ranken: „based on the inferred semantic distance of the query to the tags associated to such content, weighted by the similarity of the querying user to the users who created those tags“. Die Autoren selbst bezeichneten dies als Social Ranking. Einen vergleichbaren Ansatz gibt es von Bao et al. (2007), die Social Annotations nutzen, um die Popularität von Webseiten zu berechnen und so einen SocialPageRank zu implementieren. Die skizzierten Ansätze wurden kurz darauf von großen Websuchmaschinen wie Google oder Bing aufgegriffen, die bekannt gaben, dass sie Nutzungsdaten und Social Annotations von Facebook oder Twitter in den Retrieval- und Rankingprozess aufnehmen. Die genauen Details dieser Umsetzung bleiben im Verborgenen, aber die Grundannahme ist, dass die von Google sogenannten sozialen Kreise Einfluss auf das Ranking von Webseiten nehmen. Wird eine Webseite von einer verknüpften Person positiv markiert, so wird sie auch in dem jeweils persönlichen Ranking höher bewertet als bei anderen Personen, deren soziale Kreise diese Webseite nicht oder negativ markiert haben.

Speziell Zipfs Gesetz in Verbindung zu Folksonomien ist Bestandteil der Arbeiten von Peters (2010) und Peters und Stock (2010). Hier wird aus Zipfs Gesetz gezeigt, dass die in Folksonomien vergebenen Tags einer Zipf-typischen einer invers-logistischen oder auch einer Power-Law-Verteilung

folgen. Sie schlussfolgern, dass Tags, die im Long-Tail (also selten für ein Dokument) vorkommen, abgeschnitten werden müssen. Nur Tags mit einem hohen Wiederholungsgrad sollten für weitere Verfahren verwendet werden. Die so verbleibenden Tags seien sogenannte „Power Tags", die für ein verbessertes Retrieval in Folksonomien verwendet werden sollen. Ausgangspunkt bei Peters ist die Beobachtung, dass in Folksonomien zwar ein großes Potential im Bereich des gemeinschaftlichen Indexierens liegt, allerdings erhebliche Schwächen in der Suche und im Ranking existieren, da der Recall zwar gut sei, aber die Precision zu wünschen übrig lasse (zu „Recall" und „Presicion", vgl. Abschnitt 4.2). Peters und Stock (2010, 84) schlagen vor, Beobachtungen und Gesetzmäßigkeiten aus der Informetrie für das Information Retrieval zu nutzten. Sie unterscheiden dabei klar, wie mit den verschiedenen informetrischen Verteilungen umzugehen sei.

Peters interpretiert die Power-Law-Verteilungskurve so, dass der linke Teil der Kurve eine Wiederspiegelung der kollektiven Übereinstimmung bei der Indexierung („majorities") ist, da hier die Begriffe mit der ähnlichen Frequenz liegen („power tags"), wogegen im Tail meist Einzelmeinungen („minorities") mit niedriger Frequenz vertreten sind („tail tags"). Angewendet wird dieses Verfahren sowohl auf die Indexterme („power index tags"), als auch auf die Suchterme („power search tags"). Leider bleibt diese Arbeit eine Evaluierung des Ansatzes schuldig.

3.4.6 Sonstige Rankingverfahren

Neben den prominenten Ansätzen, wie sie in den vorherigen Abschnitten beschrieben wurden, existieren eine Reihe weniger bekannter Verfahren, auf die nur kurz eingegangen werden soll; eine ausführlichere Beschreibung ist den jeweiligen Originalquellen zu entnehmen.

In den Frühzeiten des WorldWideWeb war keines der aus der IR-Forschung bekannten Verfahren wie das Vektorraum- oder das probabilistische, sondern ein kommerzielles Modell sehr erfolgreich. Die Suchmaschine Goto.com, die später in Overtune umbenannt und letztlich von Yahoo aufgekauft wurde, setzte ein Rankingverfahren ein, das mit keinem der bisherigen Verfahren vergleichbar war. Die Position einer Webseite in der Ergebnisliste wurde (zumindest in den oberen Positionen) maßgeblich von der Höhe der Zahlungen der jeweiligen Webseitenbetreiber bestimmt (Fain und Pedersen 2006). Dies wurde unter dem Namen Sponsored Search oder Sponsored Linking bekannt und war bzw. ist eine beträchtliche Einkommensquelle für Suchmaschinenbetreiber. Zur Profitmaximierung wurde in frühen Implementationen dieses Ansatzes bewusst darauf verzichtet, bezahlte und algorithmisch bestimmte Resultate in der Ergebnisliste voneinander abzusetzen. Es war den Benutzern somit nicht klar, welches Verfahren für die jeweilige Listenpositionierung verantwortlich war.

Das Verfahren war so erfolgreich, dass die meistgesuchten Begriffe versteigert wurden, was zu hohen Summen für eine gute Platzierung in einer Trefferliste zu begehrten Begriffen wie beispielsweise „Casino“ führte. Große Suchmaschinenbetreiber wie Google oder Yahoo betreiben immer noch solche kommerziellen Rankings (z.B. Googles AdWords), doch werden erkaufte Treffer inzwischen entsprechend gekennzeichnet, bzw. räumlich von der algorithmisch erstellten Trefferliste getrennt dargestellt. Aus IR-Perspektive interessant wären Retrievaltests, die die Einschlägigkeit und Retrievalgüte von kommerziell und algorithmisch erstellen Trefferlisten vergleichen. Hierzu sind aber keine Studien bekannt.

Jordy et al. (1999) stellen einen interessanten Rankingansatz vor, der darauf basiert, die Reputation wissenschaftlicher Verlage zu messen. Das Verfahren basiert darauf, dass Buchbesprechungen dazu genutzt werden, um die Reputation eines Verlages zu schätzen. In der Studie wurden Bücher, die in den Verlagen de Gruyter, Greenwood, Doubleday, University of Georgia Press und Louisiana State University Press erschienen sind, untersucht. Die Autoren merken an, dass, obwohl die meisten Buchbesprechungen positiv formuliert sind, doch klare Variationen in der Art und Weise der Rezension erkennbar sind. Die Autoren verwenden die aus der Studie gewonnenen Daten nicht für das Retrieval. Vielmehr vergleichen sie die impliziten Reputationseinschätzungen, die aus den Besprechungen gewonnen wurden, mit denen von Bibliothekaren und prüfen, ob ein Zusammenhang zwischen Preis der Publikationen und deren Qualität bestand.

4. Evaluation von Information Retrieval-Systemen

Um die Leistungsfähigkeit verschiedener Retrievalsystemen zu evaluieren, hat sich im Bereich des IR eine spezielle Form der Evaluation etabliert, die als das Cranfield-Paradigma bekannt geworden ist. Diese Art der Laborevaluation erlaubt es, auf Grundlage von festgelegten Dokumentkorpora und Fragestellungen kontrollierte Experimente durchzuführen. Gegenüber anderen Formen der Evaluation, wie z.B. beim interaktiven Retrieval, lassen sich Cranfield-Experimente beliebig oft wiederholen. Im folgenden Kapitel wird der allgemeine Hintergrund der IR-Laborevaluation nach dem Cranfield-Paradigma und deren Möglichkeiten zur Datenanalyse erläutert.

4.1 Das Cranfield-Paradigma

Das Cranfield-Paradigma, das auf die IR-Evaluationsforschung von Cleverdon (1960) zurückgeht, ist das heutzutage beherrschende Paradigma bei der Evaluation von IR-Systemen, das einen starken Fokus auf empirische Untersuchungsmethoden legt.

Essentielle Grundlage der Evaluation sind die Testkollektionen, welche sich immer aus drei Bestandteilen zusammensetzen:

- einem Dokumentkorpus,
- einer Zusammenstellung von Informationsbedürfnissen bzw. Fragestellungen, welche als Topics bezeichnet werden und
- einer Menge an Relevanzurteilen zu den jeweiligen Topics.

Testkollektionen werden traditionell in sogenannten Evaluationskampagnen wie z.B. der Text Retrieval Conference (TREC), oder dem Cross Language Evaluation Forum (CLEF) entwickelt, gepflegt und eingesetzt. Üblicherweise werden im Rahmen der Kampagnen die Korpora und die Topics von den Organisatoren der Kampagnen zur Verfügung gestellt. Die Teilnehmer der Kampagnen entwickeln dann unabhängig voneinander ihre Retrievallösungen und reichen die Ergebnisse ihrer Systeme bei den Organisatoren ein. Hierzu wird häufig das TREC-Dateiformat als gemeinsamer Standard verwendet, das von trec_eval16 oder anderen Analyseprogrammen gelesen werden kann. Das Format ist sehr einfach aufgebaut und besteht aus:

- *Topic-ID*: Ein eindeutiger Bezeichner für das Topic. Dieser sollte aufsteigend sortiert sein. Inzwischen werden für die Topics eindeutige DOIs vergeben, so z.B. automatisch im Portal DIRECT, das in der CLEF-Kampagne eingesetzt wird.
- *Anfrageiteration*: Bezeichnet die Iteration bei mehrstufigem Retrieval. Wird in den folgenden Ad-hoc-Experimenten ignoriert.
- *Dokument-ID*: Bezeichner für das Dokument.
- *Rank*: Errechnete Rankingposition des Dokuments. Die Zählung beginnt traditionell bei 0 und die Auflistung sollte aufsteigend sein.
- *RSV-Wert*: Der vom System berechnete System-Relevanzwert. Die Auflistung geschieht aufsteigend.
- *Run-ID*: Bezeichner für das verwendete Verfahren.

Die Angabe nach diesem Standardformat erlaubt es, die einzelnen Systeme miteinander zu vergleichen. Hierzu werden die Dokumente aller Kampagnenteilnehmer in einem sogenannten Pool zusammengefasst. In diesem Pool sind alle Dokumente, die von den Teilnehmern in den trec-top-file Dateien abgeliefert wurden. Menschliche Assessoren (auch Judge oder Rater genannt) bewerten nun jedes Dokument bzgl. eines Topics hinsichtlich seiner Relevanz. Die Ergebnisse werden in Dateien festgehalten, die dem trec-rel-file Format entsprechen. Auch hier werden die Topic-ID, die Anfrageiteration (wird auch hier oft ignoriert) sowie die Dokument-ID angegeben. Zusätzlich enthält jede Zeile in der letzten Spalte noch das jeweilige Relevanzurteil des Assessors. Bei binären Relevanzurteilen sind dies die Werte 0 und 1 oder 0 bis n bei mehrstufigen

[16] <http://trec.nist.gov/trec_eval/> (Zugegriffen: 12. Oktober 2013).

Relevanzurteilen. Auf Grundlage der trec-top und der trec-rel Dateien können nun eine Reihe von Evaluationskennzahlen berechnet und das für die Kampagne leistungsfähigste System ermittelt werden.

Die zeitlich versetzte, nachträgliche Kombination von standardisierten Korpora, Topics und deren Relevanzurteilen ermöglicht eine Nachnutzung der Komponenten, die weitere Anwendungen im Rahmen der IR-Evaluation erlauben. Unabhängig von den Evaluationskampagnen kann jederzeit mit den gleichen Daten, aber einem anderen Retrievalsystem, ein anderes Experiment durchgeführt werden, das dann auf den Wissensschatz der bereits durchgeführten Relevanzbewertungen zurückgreifen kann. Dabei wird davon ausgegangen, dass ein Großteil der potentiell relevanten Dokumente in den Pools enthalten war, da unterschiedliche Kampagnenteilnehmer mit unterschiedlichen Systemen zu ihrer Erstellung beigetragen haben und so eine gewisse Streuung und Breite des Pools gewährleistet ist. Trotzdem bleibt festzuhalten, dass bei diesem Verfahren meist nicht der gesamte Korpus zu jeder Fragestellung mit Relevanzurteilen versehen ist, sondern immer nur die Dokumente, die im Pool vorhanden waren. Ein vollständig neuartiges Retrievalsystem könnte also potentiell viele nicht bewertete (aber trotzdem relevante) Dokumente finden und trotzdem schlechte Evaluationskennzahlen produzieren, da die nicht bewerteten Dokumente in der normalen Evaluation als nicht relevant eingestuft werden.

Der größte Vorteil der Cranfield-Evaluation ist die Möglichkeit, kontrollierte Experimente durchzuführen, die sich jederzeit wiederholen lassen, da alle Komponenten standardisiert und abrufbar vorliegen. Für die Sammlung von Relevanzurteilen und abgeschlossenen Evaluationen sind verschiedene Plattformen geschaffen worden, so z.B. evaluatir.org[17] oder das DIRECT-System.[18] Für Evaluationen im interaktiven Retrieval besteht diese Möglichkeit grundsätzlich nicht. Zwar können hier alle Interaktionsschritte aufgezeichnet und somit später nachgeprüft werden, aber es lässt sich nicht eine Variable des Experiments ändern und gleichzeitig die Reproduzierbarkeit des Experiments aufrechterhalten.

Auf dem IR-Forschungskontiunuum nach Kelly (2009) ist das Cranfield-Paradigma auf dem Spektrum als „systemorientierte Evaluation“ eingruppiert. Im Gegensatz zur Untersuchung von interaktiven IR und der Untersuchung menschlicher Faktoren steht das eigentliche Retrievalsystem im Vordergrund. Interaktionsabläufe oder der Kontext einer Informationssuche stehen nicht im Fokus dieses Evaluationsparadigmas. Dies führt zu kritischen Einwänden u.a. von Ingwersen und Järvelin (2005), die diese Beschränkungen überwinden wollen.

Der Vorwurf gegenüber dem Cranfield-Paradigma ist, dass die Evaluation innerhalb eines künstlich geschlossenen Systems stattfindet (Cranfield Cave),

[17] <http://wice.csse.unimelb.edu.au:15000/evalweb/ireval/about> (Zugegriffen: 12. Oktober 2013).
[18] <http://direct.dei.unipd.it> (Zugegriffen: 12. Oktober 2013).

in dem Relevanzurteile ohne den jeweiligen Kontext gefällt werden. Der Kontext, in dem sich ein Suchender befindet, sei aber für ein optimales Suchergebnis essentiell. Für diesen Ansatz, der den Anspruch erhebt, eine „holistische" Sicht auf den Retrievalprozess zu werfen, gibt es bis heute allerdings kein umfassendes Evaluationsframework, wie das des Cranfield-Paradigmas. Zwar existieren vereinzelt Insellösungen oder Testkollektionen für interaktive Retrievalevaluationen (Kelly 2009, 197-200), doch gibt es keine vergleichbare Verbreitung, wie sie die klassischen IR-Testkollektionen im TREC-Stil vorweisen können.

Nicht für alle Anwendungsbereiche des IR wird das Cranfield-Paradigma als geeignete Evaluationsumgebung angesehen, so ist z.B. beim Web-Retrieval die Größe und die Unbeständigkeit der Datenbestände ein Problem. Obwohl auch bei der Evaluation von Web-Systemen mit Assessoren und Relevanzurteilen gearbeitet wird, können die so gewonnen Daten nicht zwangsläufig für andere Experimente wiederverwendet werden. Um die Evaluierung so einfach und praxisnah wie möglich zu gestalten, gibt es hier besondere Evaluationsumgebungen, die speziell für die Evaluation von Websuchmaschinen gebaut wurden.

Um die kost- und zeitintensive Evaluation mit Assessoren zu vereinfachen, ist in den letzten Jahren der Versuch unternommen worden, mit Hilfe von Crowd-Sourcing-Systemen, wie z.B. dem Mechanical Turk von Amazon, die Evaluation durchzuführen (Alonso, Schenkel und Theobald 2010). Hierbei entstehen aber neue Probleme wie bei der Rekrutierung von geeigneten Assessoren: zum Beispiel was ihnen für Aufgaben übertragen werden können, wie man ihre Beiträge zusammenführen und wie man Missbrauch verhindern kann.

Ein anderer Trend ist die Evaluation mit Hilfe von Web-generierten Daten, z.B. die Klicks auf bestimmte Seitenelemente und Links, oder Seitenbesuche und deren Dauer zu zählen. Um die Qualität der Experimente ist es nach Crook et al. (2009) allerdings nicht immer ausreichend bestellt. So weisen die Autoren auf die Gefahren dieser Art der Experimente (z.B. A/B-Tests) hin, da zwar in vielen Wissenschaftsbereichen solche Experimente sehr verbreitet und akzeptiert seien, im Bereich der Web-Science bzw. generell „online-betriebener" Forschung aber noch nicht ausreichend entwickelt seien. Ihr Papier ist ein starkes Plädoyer für die kontrollierten Experimente der klassischen IR-Evaluation. Eine sehr ausführliche Diskussion des Cranfield-Paradigmas und der Laborevaluation im IR allgemein sowie der Vor- und Nachteile findet sich bei Sanderson (2010).

4.2 Kennzahlen der IR-Evaluation

Bei der Evaluierung von Retrievalsystemen haben sich mit fortschreitender Entwicklung dieses Wissenschaftsgebiets unterschiedlichste Verfahren und Kennzahlen zur Begutachtung der Qualität eines solchen Retrievalsystems entwickelt.

Bei der Evaluierung wird zwischen ungerankten und gerankten Ergebnislisten unterschieden. Im Folgenden werden drei Kennzahlen für Evaluierung ungerankter Ergebnislisten vorgestellt (Precision, Recall und das F-Measure) und fünf für gerankte Ergebnislisten (Cut-off-Precision, R-Precision, Average Precision, Binary Preference und Cumulative Gain). Um die jeweiligen Verfahren besser zu verstehen, beziehen sich die angebrachten Beispiele auf die erfundene Tab. 4, die von einer fiktiven, insgesamt 100 Dokumente beinhaltenden Datenbank ausgeht. Von diesen sind für eine ebenfalls fiktive Beispielanfrage 44 Dokumente relevant, wobei von diesen 30 Dokumente vom Retrievalsystem gefunden wurden und 14 nicht. Folglich sind die restlichen 56 Dokumente der Datenbank nicht für die Beispielanfrage relevant, trotzdem hat das Retrievalsystem in diesem Fall 12 davon als relevant markiert und in die Ergebnisliste mit aufgenommen. 44 nicht-relevante Dokumente wurden korrekterweise nicht vom System gefunden.

Tab. 4: Ergebnis einer Suchanfrage an eine hypothetische Datenbank mit 100 Dokumenten

	Relevant	Nicht relevant
Gefunden	30	12
Nicht gefunden	14	44

Zwei sehr oft ermittelte Werte, die zur Qualitätsbestimmung eines Retrievalsystems (aber z.B. auch von Klassifikationssystemen) herangezogen werden, sind Precision und Recall (van Rijsbergen 1974). Precision (P) ist das Verhältnis der Anzahl der gefundenen Dokumente zur Anzahl der relevanten gefundenen Dokumenten und gibt die Genauigkeit an, mit der das Retrievalsystem arbeitet:

$$\text{Precision} = \frac{|\{\text{relevante Dokumente}\}| \cap |\{\text{gefundene Dokumente}\}|}{|\{\text{gefundene Dokumente}\}|}$$

Angewendet auf die Beispieldaten aus Tab. 4 ergibt sich eine Precision von $P = 30/(30 + 12) \approx 0{,}714$, also ca. 71%.

Recall (R) ist das Verhältnis der Anzahl der relevanten Dokumente zur Anzahl der relevanten, gefundenen Dokumente. Recall kann auch als Trefferquote interpretiert werden:

$$\text{Recall} = \frac{|\{\text{relevante Dokumente}\}| \cap |\{\text{gefundene Dokumente}\}|}{|\{\text{relevante Dokumente}\}|}$$

Auch hier lässt sich der Recall mit Hilfe der Beispieldaten aus Tab. 4 berechnen. Es ergibt sich ein Recall von $R = 30/(30 + 14) = 0{,}6\overline{81}$, also ca. 68%.

Ein kombiniertes Maß aus Precision und Recall ist das sogenannte F-Measure. Hierbei wird das gewichtete, harmonische Mittel auf Precision und Recall berechnet:

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In gerankten Ergebnislisten sind die obigen Maße, die sich jeweils auf die gesamte Ergebnisliste und damit auch über alle Recall-Abstufungen erstrecken, allerdings nur bedingt einsetzbar. Um dem Nutzungsverhalten und dem konkreten Anwendungsbezug für Suchmaschinen und digitale Bibliotheken, die jeweils mit gerankten Ergebnislisten arbeiten, gerecht zu werden, wurden weitere Kennzahlen entwickelt, die Auskunft über die Retrievalqualität des Systems geben sollen. Da Nutzerstudien gezeigt haben, dass viele Benutzer eines Retrievalsystems nur die ersten Treffer einer gerankten Ergebnisliste betrachten (Hienert u.a. 2011), wird zunächst die Kennzahl Cut-off-Precision, die auch als Precision@k, $P(k)$ oder $P@k$ bekannt ist, vorgestellt:

$$P(k) = P@k = \frac{|\{\text{relevante gefundene Dokumente mit dem Rang} \leq k\}|}{k}$$

Hierbei wird nur ein Teil der vom System erzeugten Ergebnisliste zur Precisionberechnung herangezogen, bei Precision@10 (P@10) sind dies z.B. nur die ersten zehn Treffer der Ergebnisliste (Manning, Raghavan und Schütze 2008, 161). Durch das zugrundeliegende Relevanzranking wird davon ausgegangen, dass die ersten zehn Treffer – gemessen am Informationsbedürfnis des Nutzers – wahrscheinlich relevanter sind, als ein weiter am Ende der Ergebnisliste positioniertes Dokument. In den verschiedenen Evaluationskampagnen hat sich eine Abstufung der Cut-off-Precision in neun Bereiche etabliert. So sind im Evaluationswerkzeug trec_eval die neun Abstufungen P@5, P@10, P@15, P@20, P@30, P@100, P@200, P@500 und P@1000 enthalten. P@30 entspricht dabei zum Beispiel einem typischen Ergebnislistenumfang von drei Seiten (ausgehend von 10 Treffern pro Ergebnisseite). Die Berechnung der Cut-off-Precision hat den Vorteil, dass man keinerlei Abschätzung über die Gesamtzahl der relevanten Treffer benötigt; gleichzeitig lässt sie sich aber schlecht mit anderen Evaluierungsmaßen, wie z.B. dem F-Measure, verbinden. Die Gesamtzahl der relevanten Dokumente hat für eine Anfrage einen starken Einfluss auf die Cut-off-Precision.

Um dieses Problem zu umgehen, wurde die sogenannte R-Precision entwickelt. Bei dieser wird die Precision an der R-ten Position in einer gerankten Ergebnisliste, für eine Anfrage, die insgesamt $R = |\{relevante\ Dokumente\}|$ besitzt, ermittelt. Um dieses Maß bestimmen zu können, muss natürlich die Gesamtzahl der relevanten Dokumente in einem Datensatz bekannt sein. Auf den Beispieldatensatz in Tab. 4 angewendet, würde die R-Precision einer P@44 entsprechen, da es 30+14 relevante Dokumente gibt. Ein perfektes Retrievalsystem würde immer eine R-Precision von 1 erbringen, was bedeuten würde, dass alle relevanten Dokumente in der Ergebnisliste enthalten sind und ausschließlich die ersten R-Plätze der Ergebnisliste belegen würde. Die bei der

Cut-off-Precision bemängelte schlechte Vergleichbarkeit unterschiedlicher Anfragen ist mit der R-Precision besser umzusetzen.

Ein weiteres Verfahren zur Gütebewertung ist Success@n (S@n), welches die Anzahl der Anfragen angibt, die bis zu einer Trefferanzahl n zufriedenstellend beantwortet werden konnten. Konnte also beispielsweise für 12 von 20 Anfragen in den ersten 5 Treffern ein relevantes Dokument zurückgegeben werden, so ist $S@5 = 60\%$, da $(12/20 = 0{,}6 = 60\%)$.

Ein weiteres zusammenfassendes Maß ist die *Mean-Average-Precision* (MAP), das sich als sehr stabiler und aussagekräftiger Wert für die Gesamtbewertung eines Retrievalsystems herausgestellt hat. Für die Bestimmung des MAP-Wertes muss zunächst die Average Precision für jede einzelne Anfrage ermittelt werden, wobei $\mathrm{rel}(k)$ eine Indikatorfunktion ist, die den Wert 1 zurückliefert, wenn das Dokument mit Rang k relevant ist; ist es nicht relevant, liefert sie den Wert 0:

$$\mathrm{AveP} = \frac{\sum_{k=1}^{n}\big(\mathrm{P}(k) \times \mathrm{rel}(k)\big)}{|\{\text{relevante Dokumente}\}|}$$

Der MAP-Wert für eine Menge an Anfragen Q ist letztlich der Mittelwert der Average Precision jeder einzelnen Anfrage:

$$\mathrm{MAP} = \frac{\sum_{q=1}^{|Q|}\mathrm{AveP}(q)}{|Q|}$$

Speziell für den Einsatz mit bewerteter Relevanz (*graded relevance*) wurde von Järvelin und Kekäläinen (2002) *Discounted Cumulative Gain* (DCG) vorgeschlagen. Im Gegensatz zu binären Relevanzbewertungen, die nur zwischen relevant und nicht relevant unterscheiden, gibt es bei der bewerteten Relevanz weitere Abstufungen in der Relevanzbewertung. Angenommen, man verwendet eine dreistufige Relevanzskala, so sind die folgenden Relevanzbewertungen möglich: 2 – hoch relevant, 1 – teilweise relevant und 0 – nicht relevant. DCG liegt nun die Idee zugrunde, dass hoch relevante Dokumente einen höheren Einfluss auf die Berechnung der Kennzahl besitzen sollten, je höher die Rangposition ist. Ein hoch relevantes Dokument auf Rangpostition eins sollte die einflussreichste Kombination sein.

Ein immanentes Problem bei der Bewertung mittels der o. g. Kennzahlen sind unvollständig bewertete Dokumentkollektionen. Dies kommt gerade bei großen Kollektionen und bei Evaluationskampagnen häufig vor, da durch die Bewertung eines Pools anstelle der gesamten Kollektion immer nur ein Ausschnitt bewertet werden kann. Buckley und Voorhees (2004) stellen die sogenannte *Binary Preference* (bpref) vor, um dieses Problem zu umgehen. Hierbei wird ermittelt, wie oft nicht relevante Dokumente vor relevanten Dokumenten gerankt werden. Dokumente, die nicht beurteilt wurden, werden bei dieser Berechnung nicht weiter betrachtet. Die Berechnung erfolgt mittels:

$$\text{bpref} = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ höher gerankt als } r|}{R}$$

Hierbei ist n die Anzahl der nicht relevanten Dokumente und r die Anzahl der als relevant eingestuften Dokumente. R ist die Gesamtanzahl der relevanten Dokumente, analog zur Definition der R-Precision. Die bpref-Werte korrelieren meist mit MAP, zeigen aber ein stabileres Verhalten, wenn die Anzahl der bewerteten Dokumente abnimmt.

Neben den hier vorgestellten gibt es noch eine ganze Reihe weiterer Kennzahlen; da diese aber im vorliegenden Aufsatz keine Verwendung finden, wird auf eine Vorstellung verzichtet, eingehende Erklärungen finden sich in der Standardliteratur. Die hier vorgestellten Kennzahlen wurden deshalb ausgewählt, da sie die Betrachtung von jeweils unterschiedlichen Aspekten eines IR-Systems ermöglichen. Während z.B. die Cut-off-Precision den tatsächlichen Bedürfnissen eines Benutzers sehr nah kommt (möglichst viele relevante Treffer unter den ersten 10 Ergebnissen), stellt MAP einen statistisch sehr stabilen Wert dar, der sich für den Vergleich über unterschiedliche Topics und Korpora hinweg eignet. Der Einsatz von bpref ist besonders dann, wenn viele neue und unbewertete Dokumente in der Ergebnisliste vorhanden sind, von Vorteil, da sich bei der Berechnung unbewertete Dokumente auf den Wert von bpref nicht negativ auswirken. Die Auswahl der Kennzahlen stellt eine Auswahl von sowohl statistisch sehr robusten Verfahren und nutzerorientierten Werten dar.

Des Weiteren sind die o.g. Verfahren nicht frei von Widerspruch, so wird z.B. MAP von Zobel et al. (2009) als zu Recall-orientiert kritisiert, was nicht den Erwartungen der Benutzer entspräche, wobei gerade MAP von Buckley und Voorhees (2000) als besonders stabil und aussagekräftig angesehen wird. Auch die Angabe von P@n wird in manchen Kontexten als redundant eingestuft. Webber et al. (2008) vertreten die Ansicht, dass eine zusätzliche Angabe einfacher Kennzahlen wie P@n in Kombination mit komplexeren Kennzahlen wie Discounted-Cumulative-Gain keinen Mehrwert für die Interpretation der Leistungsfähigkeit eines IR-Systems bietet. Letztlich hänge dies aber immer mit dem jeweiligen Evaluationsdesign zusammen und so könne bei einer großen Anzahl an Queries eine höhere Aussagekraft mit P@n erzielt werden als mit anderen Kennzahlen. Für den Einsatz einer großen Anzahl von Queries sprechen sich auch Buckley und Voorhees (2000) aus. Eine IR-Evaluation mit weniger als 25 Queries wird als nicht aussagekräftig eingestuft, wobei 50 Queries als untere Grenze empfohlen werden, um P@n zu verwenden.

4.3 Vergleich von gerankten Listen

Beim Vergleich der Rankingergebnisse zweier Systeme ist es nicht nur wünschenswert, die Einzelleistungen der Systeme mittels der zuvor vorgestellten Kennzahlen zu quantifizieren, sondern auch, die konkreten Rankings zweier oder mehrerer Systeme miteinander vergleichen zu können. Die Frage, die hier

beantwortet werden soll, ist, inwiefern die Ergebnisse zweier Systeme miteinander korrelieren. Mathematisch kann diese Frage mittels Kendalls τ beantwortet werden (Voorhees 1998). Kendalls τ berechnet die Distanz zweier Rankings, indem die Anzahl an Vertauschungen von benachbarten Paaren gezählt wird, um ein Ranking in ein anderes zu überführen. Diese Anzahl der Vertauschungen wird normalisiert durch die Gesamtzahl von Paarvergleichen, sodass eine perfekte Korrelation den Wert 1 erhält, ein komplett gegensätzliches Ranking einen Wert von -1 und zwei vollständig zufällige Rankings einen Wert von 0. Der Rankkorrelationskoeffizient Kendalls τ ist wie folgt definiert:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$$

C ist dabei die Anzahl der Paare, die konkordant bzw. übereinstimmend sind und D die Anzahl der Paare, die diskonkordant bzw. nicht übereinstimmend sind. n ist die Anzahl der betrachteten Dokumente pro Rankingverfahren. Für folgendes, fiktives Beispiel seien die beiden Rankingverfahren R_1 und R_2 definiert, die für die Dokumente $d_1 \ldots d_5$ die folgenden Rankings liefern[19]:

$$\{d_1, d_2, d_4, d_4, d_5\} \Rightarrow$$

	d_1	d_2	d_3	d_4	d_5
R_1	1	2	3	4	5
R_2	1	3	5	3	4

Verfahren R_1 rankt die Dokumente in der Reihenfolge d_3, d_4, d_1, d_2 und d_5, Verfahren R_2 hingegen in der Reihenfolge d_3, d_4, d_2, d_5 und d_1. Um nun zu überprüfen, inwiefern die Rangordnungen übereinstimmen, müssen die Rankings R_1 und R_2 einzeln verglichen werden. Hierzu wird bei einer maximalen positiven Korrelation von R_1 und R_2 Dokument d_3 in beiden Rankings den ersten Platz belegen, d_4 den zweiten usw. Dies ist offensichtlich nicht bei allen Dokumenten der Fall. Für jedes Dokument von R_2 wird nun verglichen, ob die auf sie folgenden Rangzahlen größer oder kleiner sind. Ein konkordantes Paar liegt dann vor, wenn eine größere Rangzahl auf eine kleinere folgt, ein diskonkordantes Paar dann, wenn eine kleinere auf eine größere Rangzahl folgt (im Folgenden fett markiert).

d_3: $1-2$ $1-5$ $1-3$ $1-4$
d_4: $2-5$ $2-3$ $2-4$
d_1: $\mathbf{5-3}$ $\mathbf{5-4}$
d_2: $3-4$

Kendalls τ berechnet sich dabei nun wie folgt ($C = 8, D = 2, n = 5$):

[19] Die Notation des Beispiels geht zurück auf <http://www.univie.ac.at/ksa/elearning/cp/quantitative/quantitative-101.html> (Zugegriffen: 12. Oktober 2013).

$$\tau = \frac{8-2}{5(5-1)/2} = \frac{6}{10} = 0{,}6$$

Über die Aussagekraft der τ-Werte gibt es keine eindeutige Interpretation, wie z.B. Sanderson und Soboroff (2007) feststellen. Voorhees (2000, 712) gibt die Untergrenze von $\tau \geq 0{,}9$ für die Äquivalenz von zwei Listen an. Liefern zwei Rankings Ergebnisse, die mindestens einen solchen τ-Wert ergeben, kann von keinem signifikant unterschiedlichen Ranking gesprochen werden. Im folgenden Text wird mittels τ quantifiziert, inwieweit sich Rankings voneinander unterscheiden bzw. wie kostspielig die Überführung eines Rankings in ein anderes ist.

5. Informetrie

In den vorangegangenen Kapiteln wurde aufgezeigt, dass heutige digitale Bibliotheken in den beiden wichtigen Schlüsselkomponenten Ranking und Anfrageerweiterung Defizite aufweisen. Die üblicherweise implementierten Lösungen in diesem Bereich sind meist, gemessen an den Möglichkeiten, welche sich aus der vorhandenen Datenfülle ergeben würde, wenig innovativ. Speziell in der Gegenüberstellung von textuellen und nicht-textuellen Mehrwertdiensten zeigt sich, dass primär textuelle Dienste angeboten werden. Ein Beispiel hierfür sind Search Term Recommender, die eingesetzt werden, um die sprachliche Vagheit bei der Anfrageformulierung mit Hilfe von Thesauri auszugleichen.

Um aus der Fülle der vorhandenen (Meta-)Daten innovative Mehrwertdienste zu entwickeln, bedarf es allerdings geeigneter Verfahren. Die Informetrie als Wissenschaftsdisziplin der quantitativen Analyse von Informationseinheiten jeglicher Art bietet eine Reihe von Analyseverfahren und -modellen an, um aus den bibliografischen Daten Informationen und Wissen zu extrahieren.

> Informetrics is the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists (Tague-Sutcliffe 1992, 1).

Neben der Informetrie (als umfassender Oberbegriff) gibt es weitere Wissenschaftsdisziplinen, die sich mit der Analyse bestimmter Informationseinheiten befassen: Bibliometrie, Szientometrie und Webometrie. Diese unterschiedlichen Verfahren befassen sich jeweils mit der Messung publizierter Dokumente wie Bücher und Zeitschriften, der Messung des Produktionsergebnisses der Wissenschaft und der Messung des Informationsangebotes im Internet.

Die praktische Anwendung bibliometrischer Methoden im Bereich des Information Retrieval ereignete sich bisher selten. Laut Glänzel (2003, 9-10) sind tatsächlich nur drei praktische Anwendungsgebiete der Bibliometrie bekannt.

- *Bibliometrie für Bibliometriker*: Dies ist das Gebiet der methodischen Grundlagen-Forschung, die keinen praktischen Anwendungszweck kennt,

sondern sich mit den zugrundeliegenden Gesetzmäßigkeiten und statistischen, mathematischen oder methodischen Herangehensweisen an die Bibliometrie beschäftigt.
- *Bibliometrie für andere wissenschaftliche Disziplinen*: Die Gruppe der Wissenschaftler, die in anderen Disziplinen an bibliometrischen Fragen forscht, ist wohl die größte, aber auch die am meist durchmischte Gruppe. Bibliometrie wird hierbei natürlich mit dem Fokus auf die eigenen wissenschaftlichen Fragen und Probleme der Disziplin betrieben.
- *Bibliometrie für die Wissenschaftspolitik*: Vermehrt wird Bibliometrie zur Wissenschaftsevaluation oder als Bewilligungskriterium für Forschungsgelder oder in Berufungsverfahren verwendet. Hierbei werden meist nationale oder institutionelle Wissenschaftsstrukturen (z.B. der Einfluss deutscher Forscher im Rest Europas) in vergleichender Art betrachtet.

Zwar sind noch einige Spezialfälle (wie die Anwendung der Bibliometrie im Bereich der Wissenschaftsgeschichte) bekannt, doch stellen diese Anwendungen Ausnahmen oder Ableitungen der drei o.g. Anwendungsfelder dar. Diese Einschränkung verwundert, da doch die Bereitstellung zusätzlicher Funktionalität für das Retrieval auf Grundlage von erzeugten Metadaten von Hjørland (2000) als eine der zentralen Aufgaben von Dokumentaren und Informationswissenschaftlern angesehen wird. Ihre Funktion sei es, nicht nur die Dokumentkollektionen aufzubauen und zu unterhalten, sondern vor allem durch ihre Erschließung Mehrwerte für den Nutzer zu schaffen (Hjørland 2000, 510-2). Die zentrale Aufgabe der Informationswissenschaft sei es, allgemeines Wissen und Prinzipien für den Umgang mit Informationen zu entwickeln und durch neue Technologien nutzbar zu machen. Hjørland nennt explizit auch das Retrieval als einen solchen Mehrwert. So verwundert es, dass es offensichtlich wenige Berührungspunkte zwischen Disziplinen wie Szientometrie, Bibliometrie, Webometrie und Retrievalthemen wie IR-Systeme oder IR-Interaktion gibt.

Die für diesen HSR Focus relevanten Untersuchungsgegenstände sind bibliografische Informationseinheiten, also wissenschaftliche Papiere, ihre Autoren, das von ihnen verwendete Vokabular bzw. Fachvokabulare und Journale, Konferenzen und Verlage, in denen diese Informationseinheiten publiziert werden. Neben den gerade genannten gibt es noch die Gruppe der Referenzinformationen, die in hier aber nicht vertieft betrachtet werden. Alle genannten Informationseinheiten oder Entitäten stehen mit der zentralen Informationseinheit, dem wissenschaftlichen Papier, in Verbindung. Morris und Yen (2004) nennen auch einige der Verbindungen, die gewissen Regelmäßigkeiten oder Gesetzmäßigkeiten folgen, z.B. Lotkas Gesetz zur Autorenschaft von Papieren. Diese Gesetzmäßigkeiten und deren bibliometrische/informetrische Messung werden im folgenden Kapitel vorgestellt.

5.1 Power-Law-Verteilungen

Lange bevor der Begriff Bibliometrie geprägt wurde, publizierte Lotka (1926) seine Untersuchung zur Produktivität von Autoren. Hierzu maß er die Zahl von Einträgen in zwei Fachbibliographien – einer physikalischen und einer chemischen. Er fand wenige Autoren mit vielen Einträgen und viele, die nur ein- oder zweimal publiziert hatten. Die Verteilung der Publikationen auf Autoren war extrem schief und folgte weitgehend einer Potenzfunktion. Zu einem ganz ähnlichen Ergebnis kam wenige Jahre später der Bibliothekar Samuel Bradford (1934) bei der Untersuchung der Verteilungen von Artikeln auf Zeitschriften in den Bibliographien zweier Spezialgebiete: Wenige Kernzeitschriften enthielten den größten Teil der Literatur, während eine ganze Reihe von Zeitschriften im betrachteten Zeitraum jeweils nur einen Aufsatz zum Thema herausbrachten. Dies ist der wesentliche Inhalt des nach ihm benannten Gesetzes der Streuung von Literatur (Bradford's Law of Scattering).

Beobachtungen aus der Bibliometrie, der Sprachnutzung oder der Wirtschaft zeigen, dass viele Verteilungs- und Entwicklungsphänomene – seien sie natürlichen oder menschlichen Ursprungs – einem Power-Law (dt. Potenzgesetz) entsprechen (Newman 2005). Allgemein lässt sich diese Art der Verteilung mit dem folgenden Ausdruck beschreiben:

$$f(x) = cx^{-\alpha},$$

wobei c ein konstanter, normalisierender Faktor, x der Rang der aktuellen Informationseinheit und α die Stärke des Gefälles ist. Power-Law-Funktionen sind monoton, d. h. wenn x seinen Wert verändert, wird $f(x)$ stetig ansteigen oder abfallen. Wenn Power-Law-Funktionen genutzt werden, um Verteilungen zu beschreiben, ist der Exponent α typischerweise positiv. Dies führt dazu, dass, wenn x steigt, $f(x)$ fällt. Verallgemeinert bedeutet dies, dass gezählte Items oder Ereignisse mit hohem Aufkommen typischerweise selten vorkommen. Man spricht bei den typischen Verteilungen auch von einem sogenannten Long-Tail oder der 80:20 Regel. Power-Law-Verteilungen sind unter verschiedenen Namen in der Literatur bekannt. In den folgenden Abschnitten sollen daher die Pareto-Verteilung und die Gesetzte nach Lotka, Zipf und Bradford vorgestellt werden.

5.2 Pareto-Verteilung

Das nach Vifredo Pareto (1848-1923) benannte Gesetz der Pareto-Verteilung lautet frei zusammengefasst: 80% aller erkennbaren Effekte resultieren aus 20% aller Ursachen. Zunächst ist festzuhalten, dass Pareto diese Gesetzmäßigkeit nie selbst unter diesem Namen verbreitete. Der Begriff „Pareto's Law" wurde vielmehr von Juran (1954) geprägt. Pareto selbst wird nachgesagt in seiner Arbeit nachgewiesen zu haben, dass 80% des Eigentums in Italien in der Hand von 20% der Italiener läge. In zahlreichen Arbeiten von anderen Autoren

wurde u.a. beobachtet, dass 20% der Angestellten eines Unternehmens 80% der Produktivität erwirtschaften, dass 20% der Kunden 80% des Umsatzes erzeugen, dass 20% eines Bibliotheksbestandes 80% der Ausleihen erzeugen, dass 80% der Weltbevölkerung in 20% der Städte leben usw. Man spricht daher umgangssprachlich auch von der 80:20 Regel.

Pareto selbst betrachtete mehr als 40 Datensätze zur Einkommensverteilung in Europa, Nord- und Südamerika und leitete daraus folgenden Zusammenhang ab:

$$\ln N = \ln A - \alpha \ln x.$$

Hierbei ist N die Anzahl der Personen, deren Einkommen größer als x ist. A und $\alpha > 0$ sind Parameter, die geografisch und zeitlich variieren.

Ein Beispiel soll dieses Modell erläutern. Tab. 5 zeigt einen Auszug aus Paretos Daten zu britischen Steuerzahlern aus den Jahren 1893/1894. Der Wert x gibt dabei die Einkommensgruppe und N die Anzahl von Steuerzahlern an, die mindestens ein entsprechendes Einkommen hatten. So gab es 22.896 von insgesamt 400.648 Steuerzahlern, die mindestens 1.000 £ zu versteuern hatten. Es kann nun berechnet werden, wie viele Personen ein Einkommen z.B. von 2.500 £ zu versteuern hatten:

$$\exp(19{,}331 - 1{,}3379 \ln 2500) \approx 7067.$$

Die beiden Parameter $A = 19{,}331$ und $\alpha = 1{,}3379$ wurden dabei aus Hardy (2010) entnommen.

Tab. 5: Paretos Datensätze zur Einkommensverteilung (nach Hardy, 2010)

x	N
150	400.648
200	234.185
300	121.996
400	74.041
500	54.419
600	42.072
700	34.269
800	29.314
900	25.033
1.000	22.896
2.000	9.880
3.000	6.069
4.000	4.161
5.000	3.081
10.000	1.104

Der Wert x beschreibt hierbei das Einkommen in £, N ist die Anzahl der Steuerzahler Großbritanniens im Jahr 1893/1894, die in die entsprechende Einkommensgruppe fallen.

5.3 Lotkas Gesetz

1926 veröffentliche Alfred J. Lotka sein später oft zitiertes Papier mit dem Namen „The frequency distribution of scientific productivity“, in dem er die später nach ihm benannte Gesetzmäßigkeit formulierte:

> [...] the number (of authors) making n contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 per cent (Lotka 1926).

Eine beispielhafte Rechnung, bei der 100 Autoren jeweils einen Artikel verfassen, würde ergeben, dass es 25 Autoren gibt, die je 2 Artikel veröffentlichen $(100/2^2 = 25)$. 11 Autoren hätten jeweils 3 Papiere verfasst $(100/3^2 \approx 11)$ usw. Lotka selbst spricht hierbei von einer inversen Quadrat-Gesetzmäßigkeit (*inverse square law*). Seine These leitete er aus der Beobachtung zweier bibliografischer Datensätze aus der Chemie und Physik ab (Chemical Abstracts und Geschichtstafeln der Physik). Die prozentualen Anteile der Autoren, die $1, 2, 3, \ldots\ n$ Papiere veröffentlicht hatten, plottete er in einen Graphen mit doppelt logarithmierten Skalen. Den Grad der Steigung (*slope*) bestimmte er mit der Methode der kleinsten Quadrate und kam dabei auf eine negative Steigung von ca. 2. Lotka hatte damit eine Gesetzmäßigkeit über die Produktivität von Autoren im Wissenschaftsbetrieb aufgestellt.

Bis 1941 wurde die Arbeit zunächst nicht zitiert und wurde erst ab 1949 unter dem Namen „Lotka's Law“ bekannt (durch Zipf, 1949). Spätestens durch die Arbeiten von de Solla Price (1963) wurde sie allerdings sehr populär. Auch wenn viele Autoren, darunter auch de Solla Price, behaupteten, dass Lotkas Beobachtungen für unterschiedlichste Gebiete der Wissenschaft und deren Produktivität gültig seien, kann dies zunächst auf Grundlage zu geringer Größe der Datensätze und auf Grund statistischer Unzulänglichkeiten nicht mit Sicherheit behauptet werden. Wie Potter (1981, S. 22 ff.) zeigt, sind viele frühe Untersuchungen zu Lotkas Gesetz nicht in der Lage, einen statistischen Signifikanztest nach Kolmogoro-Smirnov zu bestehen. Hierbei wird die maximale Abweichung (*deviation*) D ermittelt:

$$D = \max\left|F_0\left(X - S_n(X)\right)\right|.$$

Potter beschreibt $F_o(X)$ als die theoretische kumulative Frequenzfunktion und $S_n(X)$ als die tatsächlich beobachtete kumulative Frequenzfunktion eines Datensamples aus n Beobachtungen. Signifikanz sei bis zu einem Schwellenwert von 0,01 gegeben, wobei die Kolmogoro-Smirnov-Statistik gleich $1{,}63/n^2$ ist. Ist D größer als die Kolmogoro-Smirnov-Statistik, dann ist die Beispielverteilung nicht gleich der theoretisch angenommenen Verteilung. Lotka selbst betrachtete für die Chemical Abstracts 6.981 Autoren, sodass Kolmogoro-Smirnov hier den Wert $1{,}63/\sqrt{6981} = 0{,}0195$ annimmt. Für die Geschichtstafeln der Physik liegt der Wert bei $1{,}63/\sqrt{1325} = 0{,}0448$. Laut Potter erfüllen aber nur die ermittelten Werte für D der Geschichtstafeln der Physik die o.g. Bedingung; für die Chemical Abstracts könne Lotka selbst seine Gesetzmäßig-

keit nicht einhalten (s. Tab. 6). Gleiches gelte auch für viele andere Studien, die bis in die 1980er Jahre durchgeführt wurden und die nur teilweise Lotkas Beobachtungen voll erfüllen konnten. Trotzdem ist in Tab. 6 sehr gut zu sehen, dass für kleine Werte (Anzahl der veröffentlichten Papiere pro Autor) eine größere Schwankung zwischen beobachteten und erwarteten Werten zu sehen ist als für große Werte.

Tab. 6: Lotkas Daten der Chemical Abtracts

Anz. der Papiere	Beobachtet	$S_n(X)$	Erwartet	$F_0(X)$	$\lvert F_0(X - S_n(X))\rvert$
1	0,5792	0,5792	0,6079	0,6079	0,0287
2	0,1537	0,7329	0,1520	0,7599	0,0270
3	0,0715	0,8044	0,0675	0,8274	0,0230
4	0,0416	0,8460	0,0380	0,8654	0,0194
5	0,0267	0,8727	0,0243	0,8897	0,0170
6	0,0190	0,8917	0,0169	0,9066	0,0149
7	0,0164	0,9081	0,0124	0,9190	0,0109
8	0,0123	0,9204	0,0095	0,9285	0,0081
9	0,0093	0,9297	0,0075	0,9360	0,0063
10	0,0094	0,9391	0,0061	0,9421	0,0030

Da $D > 0{,}0195$ folgen die Daten streng genommen nicht Lotkas Gesetz. Tabellenwerte übernommen von Potter (1981).

Die Gründe hierfür sind vielfältig. Zunächst ist, wie bei allen bibliografischen Modellen, Lotkas Gesetz keine präzise statistische Verteilung, sondern eine Generalisierung, die auf zwei Datensätzen beruht. Weiterhin wurde von anderen Autoren der Fehler begangen, zu kleine Datensätze zu betrachten, die eine schlechte thematische oder zeitliche Abdeckung aufwiesen. Weiterhin wurden teilweise nur einzelne Quellen (u.a nur eine einzige Zeitschrift) betrachtet, was die Datenlage zusätzlich verzerrt. Darüber hinaus untersuchte Lotka selbst in seinem Datensatz nur die Autoren mit den Anfangsbuchstaben A und B, weiterhin betrachtete er nur die Erstautoren eines Papiers, die Koautoren wurden von ihm ignoriert (Havemann 2009).

Das Problem der Produktivitätsmessung durch Publikationszahl ist nicht unumstritten und im Grunde das häufigste Argument gegen den Science Citation Index (SCI). Umstätter (1999) nennt fünf Gründe, die hinter einer hohen Publikationsrate stehen können:

1) Der Autor beschreibt eine Methode, die sich auf viele Probleme anwenden lässt. Papiere solcher Art werden laut Umstätter allerdings auch sehr häufig zitiert.
2) Der Autor beschreibt eine Theorie, die sich auf viele Probleme anwenden lässt.
3) Der Autor beschreibt eine umstrittene Behauptung, die hohe Attraktion für Zeitschriften und Leser hat.
4) Der Autor besitzt einen hohen Bekanntheitsgrad bzw. eine hohe berufliche Position.

5) Der Autor besitzt eine überdurchschnittliche Publikations- und Einreichungsaktivität. Er reicht seine zahlreichen Publikationen so oft bei unterschiedlichen Zeitschriften und Verlagen ein, bis letztlich das Papier veröffentlicht wird.

Trotz der Schwächen bleibt Lotkas Gesetz ein typisches und höchstprominentes Beispiel für eine sogenannte Size-Frequency-Verteilung (Egghe 2005), welche die Anzahl der Quellen (Sources) mit einer bestimmten Nummer von Erzeugnissen (Items) in Verbindung setzt. Laut Potter (1981) ist es gleichzeitig ein starker Hinweis auf die Existenz universeller Gruppen von Autoren, die nach ähnlichen Produktivitätskriterien publizieren. Je größer und thematisch, zeitlich und international umfassender eine solche Gruppe wird, umso mehr scheint sie Lotkas Gesetz zu folgen, wohingegen kleinere Forschergruppen (z.B. einer bestimmten Unterdisziplin oder einer kurzen zeitlichen Epoche) mitunter nach anderen Regelmäßigkeiten publizieren. Er hält außerdem fest, dass, um Lotka-typisches Verhalten nachweisen zu können, ein Zeitraum von mindestens zehn Jahren betrachtet werden sollte. Eine weitere Schwierigkeit sind die zuvor beschriebenen Probleme bei der eindeutigen Identifikation von Autoren in digitalen Bibliotheken oder Bibliografien. Die eindeutige Identifizierung wird z.B. durch nicht einheitliche Erschließung der Namen oder durch mehrdeutige Namen erschwert. In großen Datenbeständen kann eine Analyse auch ohne eine eindeutige Identifizierung der Autoren möglich sein, z.B. durch Hilfsmittel wie die Detektion von Kohorten oder Langzeitautoren.

5.4 Zipfs Gesetz

> The Principle of Least Effort means ... that a person ... will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems (Zipf 1949).

Obwohl dieses Zitat auf den ersten Blick zunächst wenig mit Informetrie zu tun haben scheint, ist es doch einer der einflussreichsten Grundsätze der Informetrie. Die von Zipf aufgestellte Gesetzmäßigkeit der Wortverteilungen in der natürlichen Sprache besagt, dass das Produkt aus Rang r und Frequenz f konstant C sein sollte. Die von Zipf aufgestellte Formel lautet

$$C = r \cdot f.$$

Zipf sortierte alle Wörter eines Textes nach Häufigkeit und trug auf der Ordinatenachse die jeweilige Häufigkeit auf. Auf die Abszissenachse trug er die Rangreihenfolge auf und erhielt so eine Hyperbel. In Tabelle 7 ist ein Beispiel für Zipfs Untersuchungen aufzeigt, welches den Rang und die Häufigkeit unterschiedlicher Wörter in der Ulysses von James Joyce auflistet. Das zehnthäufigste Wort kommt danach 2.653 mal vor, das 20.000 häufigste allerdings nur ein einziges Mal. Bei diesem Beispiel ist $C \approx 24.500$.

Tab. 7: Die Häufigkeitsverteilung aller Wörter aus Ulysses von James Joyce als Beispiel für Zipfs Gesetz

r	f	$C = r \cdot f$
10	2.653	26.530
20	1.311	26.220
30	926	27.780
40	717	28.680
50	556	27.800
100	265	26.500
200	133	26.600
300	84	25.200
400	62	24.800
500	50	25.000
1.000	26	26.000
2.000	12	24.000
3.000	8	24.000
4.000	6	24.000
5.000	5	25.000
10.000	2	20.000
20.000	1	20.000
29.899	1	29.899

Zu sehen sind Rang r, Häufigkeit f und das nahezu konstante Produkt aus r und f. Entnommen aus Glänzel (2003, 42).

Nach Umstätter (2003) sprach Zipf bei den beiden Achsen auch von Kräften: Die Ordinatenachse nannte er „force of unification“, die Abszissenachse „force of diversification“. Die erste Kraft führt dazu, dass bestimmte Wörter im Englischen wie „the“, „it“, „is“ usw. oft vorkommen. Umstätter rechnet, dass jedes 10. bis 20. Wort in einem englischen Text ein „the“ ist. Gleichzeitig werden aber auch viele Wörter benutzt, die im gesamten restlichen Text nicht mehr oder nur sehr selten vorkommen. Prinzipiell ist diese Gesetzmäßigkeit für alle Sprachen gegeben, dabei allerdings in unterschiedlichen Ausprägungen. So können im Deutschen die Äquivalente zum englischen Artikel „the“, „der“, „die“ und „das“ als genus-markierende Artikel aufgrund ihrer anderen Verteilung nur zu einem Drittel so häufig vorkommen.

Das Principle of Least Effort ist allerdings für jeden Text-erzeugenden Menschen das Gleiche. Es ist, wie von Shannon und Weaver beschrieben, aus der Sicht der Informationstheorie nichts anderes, als ein ausgewogenes Verhältnis von Information und Redundanz. Wendet man ein künstlich verkleinertes oder kontrolliertes Vokabular an, wie einen Thesaurus oder eine Klassifikation, kann dies starke Auswirkungen auf die force of unification bzw. die force of diversification haben, da somit bewusst mit einem natürlichen Spracheinsatz gebrochen wird. Vergleicht man zum Beispiel das Sprachlernverhalten von Kindern, so sieht man, dass dieses fast natürlich dem Zipfschen Gesetz folgt: Ein Kind lernt pro Tag ca. drei Wörter und je nach Bildungsgrad später bis zu 10.000 insgesamt. Die Auswahl erfolgt dabei zum einen nach der Häufigkeit

des Gebrauchs (permanente Wiederholung suggeriert Wichtigkeit) oder nach Neuigkeit und damit einem besonders hohen Informationsgehalt. So ergibt sich auch Zipfs Ansatz, dass der Zusammenhang zwischen Information · Redundanz = Konstante lautet.

5.5 Bradfords Gesetz

Das sogenannte Bradfordsche Gesetz oder Bradfords's Law of Scattering wurde als Begriff von Brian C. Vickery geprägt. Bradford selbst beschrieb seine Beobachtung wie folgt:

> [...] if scientific journals are arranged in order of decreasing productivity on a given subject, they may be divided into a nucleus of journals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus when the number of periodicals in the nucleus and the succeeding zones will be as $1: b: b^2$ ... (Bradford 1934).

Die Grundaussage von Bradfords Gesetz lautet, dass sich eine immer gleiche Menge von Literatur auf gerankte Zeitschriften nach dem Prinzip $1: b: b^2$ ... bzw. $n^0: n^1: n^2$ ... verteilt. Ein einfaches Beispiel macht diesen Sachverhalt deutlich: Während die Journale „Nature“ oder „Science“ von vielen Naturwissenschaftlern häufig gelesen werden, gibt es zahllose Zeitschriften, die nur wenige Leser mit einem entsprechenden Spezialinteresse haben. Gleiches gilt in umgekehrter Reihenfolge auch für die Zusammensetzung der Zeitschriften, die von thematisch sehr breit bis hochspeziell reichen. Anders ausgedrückt heißt dies, dass ein Kern (der Nukleus) von wenigen Zeitschriften einen großen Teil der für ein bestimmtes Thema relevanten Literatur enthält (die sogenannten Kernzeitschriften), wobei hingegen eine große Zahl von Zeitschriften existiert, die nur sporadisch zu einer bestimmten Thematik publizieren. Möchte man nun eine vollständige Bibliographie zu einer Thematik erstellen, muss man einen großen organisatorischen, logistischen wie finanziellen Aufwand betreiben, um die Thematik abzudecken (Havemann 2009, 16).

Bradford selbst führte seine Analysen aus, indem er die Zeitschriften in der Rangfolge entsprechend der Anzahl der zur Thematik relevanten Artikel ordnete und jeweils gleich große Gruppen bzw. Zonen bildete. Der sogenannte Nukleus besteht hierbei aus 10 Zeitschriften, die insgesamt ein Drittel der 398 relevanten Artikel zur Schmiermittel-Forschung (Lubrication Research) beinhalten. Die zweite Zone beinhaltet 35 weitere Zeitschriften und die dritte Zone ca. 122. Jede Zone beinhaltet folglich ungefähr das 3,5-fache der vorherigen Zone. Auch hier würde man bei einer anderen Darstellungsweise, die zunächst nicht auf kumulierte aber gleichzeitig doppelt-logarithmierte Achsen setzt, eine für Power-Laws typische Gerade erkennen. Die Einteilung in drei Zonen ist rein willkürlich. Je nachdem, wie viele Dokumente einer Zone zugerechnet werden, können beliebig viele Zonen definiert werden.

5.6 Lotkaische Informetrie und der Informationsproduktionsprozess

Alle zuvor beschriebenen Beobachtungen folgen einer einfachen Regel, die nicht mit normalen und bekannten Gauß- oder Poisson-Verteilungen zu vergleichen ist. Allgemein sind dies Phänomene, die ein starkes Ungleichgewicht in den Verteilungen erkennen lassen. Alle zuvor beschriebenen Gesetzmäßigkeiten – seien es Lotkas, Zipfs oder Bradfords Gesetz – beschreiben letztlich das gleiche Phänomen.

Mathematisch wurde dieser Zusammenhang sowie eine Überführung der unterschiedlichen Notationen z.B. von Adamic (2000) dargestellt. Adamic zeigt diesen Zusammenhang mit Hilfe eines AOL-Datensatzes von Log-Dateien. Sie zeigt dabei auch, dass die lange bekannten bibliometrischen Gesetzmäßigkeiten auch im WWW wiederzufinden sind. So zeigen sich die gleichen o.g. Phänomene in der Anzahl der Besucher einer Webseite, die Anzahl der Webpage innerhalb einer Website oder die Anzahl der Links zu einer Webseite. Auch neuere Arbeiten zeigen immer wieder, dass es sich bei all diesen Beobachtungen, seien sie dem Bereich der Bibliometrie, Webometrie oder allgemein der Informetrie zuzuordnen, letztlich um gemeinsame Regelmäßigkeiten handelt. Bei der mathematischen Herleitung wird die Äquivalenz deutlich und die zugrundeliegenden Power-Laws werden als gemeinsames Konstrukt sichtbar.

Die Gemeinsamkeiten der Gesetzmäßigkeiten werden von Egghe (2005) nicht nur mathematisch hergeleitet, sondern auch konzeptionell untermauert. Wird in den jeweiligen Gesetzen jeweils von einer konkreten Entität wie bspw. Zeitschriften, Zeitschriftenartikeln, Autoren oder Quellen gesprochen, schlägt Egghe vor, von Sources (für Zeitschriften, Autoren etc.) und *Items* (welche von den *Sources* produziert werden, z.B. Artikel) zu sprechen. Ganz allgemein fasst er diese Art der mathematischen Analyse der o.g. Phänomene als lotkaische Informetrie (*Lotkaian Informetrics*), und den Prozess, der zu solchen Phänomenen führt, als Informationsproduktionsprozess *(Information Production Process*, IPP) zusammen. Beispiele für Sources und Items sind Autoren, die Artikel schreiben; Bücher in Bibliotheken, die Ausleihen erzeugen; oder Wörter, die in Texten Verwendung finden (s. Tab. 8).

Die dem IPP zugrundeliegende Source/Item-Beziehung kann auch als Type/Token-Beziehung bezeichnet werden. Es sollte dabei unterschieden werden zwischen der Size-Frequency-Funktion f (typischerweise Lotka) und der Rank-Frequency-Funktion g (typischerweise Zipf). Während bei der Size-Frequency-Funktion für jedes Anzahl $n \in \mathbb{N}$ von Items die Anzahl $f(n)$ der Sources bestimmt werden kann, ist auch gleichzeitig für den Rang r einer Source $s \in S$ die Funktion $g(r)$ definiert, die die Anzahl der Items, die zu einer Source gehören, liefert. Beide Funktionen $f(n)$ und $g(r)$ sind dual (Egghe 2005, 10-2). Diese Dualität im IPP bedeutet, dass prinzipiell die Gesetzmäßigkeiten von Sources mit Hinblick auf ihre jeweiligen Items als auch die Gesetzmäßigkeiten von Items mit Hinblick auf deren Sources betrachten werden

können. Sources und Items sind austauschbar. Die Size-Frequency-Funktion $f(n)$ liefert die Anzahl der Sources mit $n \in \mathbb{N}$ Items, wobei die Rank-Frequency-Funktion $g(r)$ die Anzahl der Items in der Source mit dem Rang r liefert. Demnach würde dies für die Artikel/Referenzen einmal eine aktive „zitiert" sowie eine passive „wird-zitiert"-Beziehung bezeichnen. Es ist folglich wichtig, die Art der Source/Item-Beziehung explizit anzugeben, um Verwirrungen vorzubeugen. Prominente Beispiele für solche Source/Item-Beziehungen sind in Tab. 8 aufgelistet. Egghe verwendet in seinen Arbeiten den Parameter n für $f(n)$; da in diesem Beitrag jedoch allgemeiner von Power-Laws gesprochen wird, wird hier die Bezeichnung $f(x)$ synonym verwendet.

Tab. 8: Beispiele für Source/Item Beziehungen in einem IPP, angelehnt an Egghe (2009, 2)

Sources		Items
Autoren	→	Artikel
Zeitschriften	→	Artikel
Artikel	→	Zitationen (von/zu)
Artikel	→	Koautoren
Bücher	→	Ausleihvorgänge
Wörter (Typen)	→	Nutzung von Wörtern in Texten (Token)
Websites	→	Hyperlinks (ein-/ausgehend)
Websites	→	Webpages
Städte/Dörfer	→	Einwohner
Angestellte	→	Ihre Produktivität
Angestellte	→	Ihr Gehalt

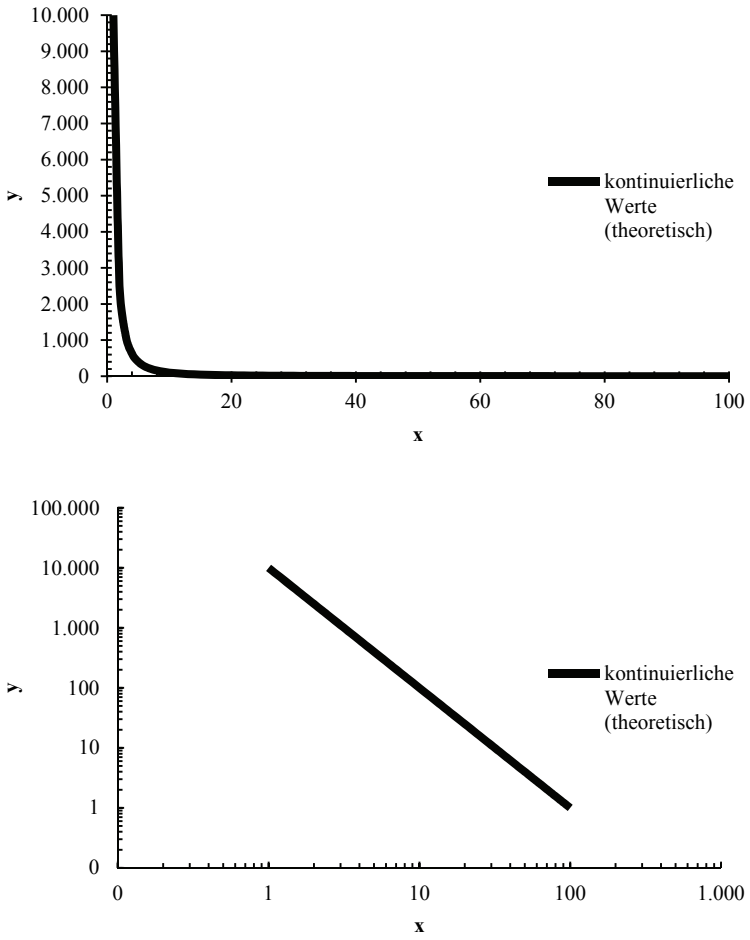
5.7 Eigenschaften von Power-Laws

Nachdem nun die Power-Laws als charakteristisches Merkmal hinter der Lotkaischen Informetrie und den IPPs aufgezeigt wurden, sollen im folgenden Abschnitt einige Eigenschaften der Power-Laws beschrieben werden. Zunächst werden diese auch als skalenfrei (scale-free) bezeichnet. Dies sagt aus, dass eine Erhöhung um einen bestimmten Faktor für den Wert von x eine entsprechende Verringerung (oder Erhöhung) für $f(x)$ bedeutet. Allerdings hat Newman (2005) nachgewiesen, dass die Power-Law-Verteilung die einzige skalenfreie Verteilung ist, sodass die beiden Bezeichnungen synonym verwendet werden können. Im Rahmen dieses HSR Focus wird allerdings die Begrifflichkeit der Power-Laws bevorzugt verwendet.

Abb. 1 zeigt zwei typische Visualisierungen für eine lehrbuchhafte Power-Law-Funktion ($c = 10.000;\ a = 2$). Zunächst ist die charakteristische, stark abfallende Kurve in der ersten Visualisierung mit nicht-skalierten Achsen zu sehen. Man erkennt sehr gut den Long-Tail, der gegen die y-Achse konvergiert. Die gleiche Funktion kann auch mit einer doppelt logarithmierten Skala gezeichnet werden, was durchaus üblich ist. In dieser Ansicht erkennt man eine Gerade, was als ein erstes Indiz für eine Power-Law-Verteilung der gemesse-

nen Daten angesehen werden kann. Später wird neben der grafischen Identifikation auch ein mathematisches Verfahren vorgestellt, um Power-Law-Verteilungen zu erkennen und den Exponenten α zu errechnen.

Abb. 1: Plot derselben Power-Law-Funktion mit den Werten $c = 10.000$ und $\alpha = 2$.



Während der erste Plot eine natürliche, nicht-skalierte Achse besitzt, wurden im zweiten Plot beide Achsen logarithmisch skaliert. Die dargestellte Funktion ist ein typischer Vertreter für ein theoretisches Power-Law, das auf kontinuierlichen Werten basiert.

5.8 Vergleich von empirisch ermittelten und formalen Power-Laws

Vergleicht man empirische Power-Law-Verteilungen, also gemessene Werte wie Stadtpopulationen oder Telefonanrufe (vgl. die vorherigen Abschnitte), mit mathematischen bzw. formalen Power-Laws, so wird schnell klar, dass empirisch hergeleitete Power-Law-Verteilungen nur eine Approximation ihrer mathematischen Beschreibungen sind. Durch die zwangsläufige Vereinfachung eines umfangreichen Sachverhaltes in die Form eines mathematisch simplen Power-Law-Modells ist eine Abweichung nicht zu vermeiden. Die durch Modelle beschriebenen Sachverhalte sind meist komplexerer Natur oder ihre Umstände mathematisch schwieriger zu beschreiben. Dies gilt auch für die in dieser Arbeit betrachteten Anwendungsfälle der empirisch gemessenen Verteilungen im IPP. Wie in allen Modellbildungsprozessen nimmt man mit der Approximation eine Abweichung von der wirklichen Verteilung in Kauf. Können einfache Power-Laws die gemessenen Datenlage nicht ausreichend genug wiedergeben oder die Repräsentation eine systematische Abweichung aufweisen, sollte eine Anpassung des einfachen Power-Law-Modells in Betracht gezogen werden. Milojević (2010) fasst insgesamt vier solcher Abweichungen zusammen, die nachfolgend beschrieben werden.

Zunächst beschreiben Csányi und Szendrodblaci (2004, 036131-2) eine typische Abweichung, die sich durch die Existenz zweier unterschiedlicher Exponenten für die klassische Power-Law Formel auszeichnen. In doppelt logarithmierter Darstellungen ist nicht eine, sondern sind zwei Geraden zu sehen.

In Glänzel (2007, 94) wird die Pareto-Verteilung zweiter Ordnung („Pareto distribution of the second kind“) oder die sogenannte Lomax-Verteilung vorgestellt, welche vorliegt, wenn für eine positive Zufallsvariable X gilt:

$$G(x) = P(X \geq x) = \frac{N^{\alpha}}{(N+x)^{\alpha}}, \text{ für alle } x \geq 0,$$

wobei N und α positive reelle Zahlen sind. Alternativ wird laut Glänzel auch die Parametertransformation $\alpha = \alpha/(\alpha - 1)$ angewendet. Praktisch bedeutet dies, dass die Lomax-Verteilung zunächst für kleine Werte von x nahezu konstant erscheint (im Plot erkennt man eine fast horizontale, schwach abfallende Linie), wobei für große Werte von x eine reguläre Power-Law-Verteilung sichtbar ist.

Eine weitere Variante sind die in Newman (2005) vorgestellten Power-Laws mit einem exponentiellen Cut-off. Hier wird für kleine Werte von x eine normale Power-Law-Verteilung beobachtet, sodass im Plot zunächst eine gerade Linie zu erkennen ist. Für größere Werte von x fällt die Gerade allerdings ab und ähnelt dem Plot einer exponentiellen Funktion. Der exponentielle Anhang für große x fällt also schneller als das Power-Law.

Eine von Milojević (2010) selbst beschriebene Variante sind die der Log-Normal/Power-Law-Verteilungen, bei denen für große x eine normale Power-Law-Verteilung zu sehen ist, für kleine x allerdings eine logarithmische Nor-

malverteilung vorliegt. Diese Normalverteilung zeigt sich im doppelt logarithmierten Plot durch eine Rundung, die zusätzlich ein Maximum besitzen kann – gegensätzlich zu normalen Power-Law-Verteilungen oder den drei vorgestellten Varianten, die kein solches Verhalten zeigen und alle monoton fallend sind.

Neben den genannten vier Variationen gibt es einen weiteren Faktor, der einen Unterschied zwischen formalen und empirisch ermittelten Verteilungen erklärt. Die formale Definition geht von einem kontinuierlichen Zahlenraum für x aus; tatsächlich gemessen werden können aber nur diskrete Werte $x \in \mathbb{N}$, die zu einer Verzerrung in der Darstellung führen. Diese Verzerrungen sind im Vergleich der Abb. 1 für theoretische, kontinuierliche Werte und Abb. 2 für empirisch ermittelte diskrete Werte zu sehen. Bei theoretisch, künstlich generierten Werten ist der sogenannte Fat-Tail zu sehen, der auf dem Phänomen beruht, dass es wesentlich mehr Werte zum Ende der Verteilung gibt. Die gedachte Linie ist aber sowohl hier, als auch bei empirisch gemessenen diskreten Werten klar zu erkennen. Bei den abgebildeten diskreten Werten handelt es sich um einen Plot der Verlagsinformationen aus der Datenbank Bibsonomy (verwendet wurde der Bibsonomy-Dump-2011-01-01). Klar zu erkennen ist sowohl der Fat-Tail als auch das für empirisch ermittelte Verteilungen typische Zittern im Tail.

Es ist üblich, für das kontinuierliche Modell die Variable x und für die diskrete Variante die Variable k zu verwenden, wie dies auch in den Abbildungen an den Achsenbeschriftungen zu sehen ist. Weiterhin kann gezeigt werden (Egghe 2005, 378-80), dass die kontinuierliche Funktion

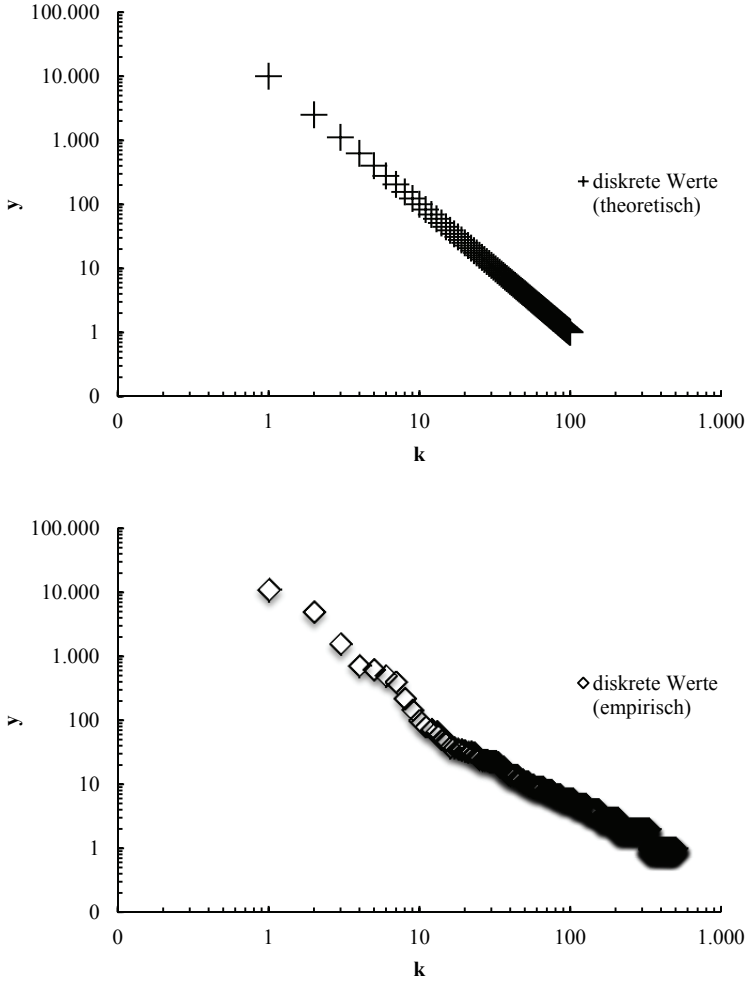
$$f(j) = \frac{C}{j^{\alpha}},$$

die gleichbedeutend mit der klassichen Power-Law-Formel ist, für $j \in [1, \rho_m]$ für $\rho_m = \infty$ mathematisch auf die folgende diskrete Form abgebildet werden kann:

$$f_d(n) = \frac{K}{n^{a}}.$$

Hierbei ist $n = 1, 2, \ldots n_{max}$ und K konstant. Man beachte die unterschiedliche Benennung des Exponenten, die α für die kontinuierliche Funktion und a für die diskrete Variante lautet. Diese Unterscheidung stammt aus der Arbeit von Egghe (2005), sie ist aber in anderen Arbeiten nicht unbedingt üblich, sodass in hier sowohl für die kontinuierliche als auch die diskrete Variante die Variante α verwendet wird.

Abb. 2: Plot zweier diskreter Power-Law-Verteilungen



Basierend auf einer theoretisch und einer empirisch ermittelten Werteverteilung zu Verlagsinformationen in der Datenbank Bibsonomy.

Eine weitere Diskrepanz ist anhand der Beispielrechnung zu Lotkas Gesetz und einer an Milojević (2010) angelehnten Rechnung zu erklären. Für den Exponenten $\alpha = 2$ und dem normalisierenden Faktor $c = 100$ würden 6,25 Autoren vier Papiere verfasst haben ($f(4) = 100 * 4^{-2} = 100/4^2 = 6{,}25$). In einem wirklichen Beispiel muss der Wert 6,25 auf eine natürliche Zahl abgebildet werden, üblicherweise wird auf den Wert 6 abgerundet. Die Rundungsproble-

matik kann aber nicht den wahren Kern des Problems beschreiben, da empirisch ermittelte Power-Laws nicht einfach nur gerundete Varianten eines theoretischen Power-Laws sind. Vielmehr muss die probabilistische Eigenart der empirischen Verteilungen betrachtet werden, welche besagt, dass der ermittelte Wert von 6,25 nur den Mittelwert aller Verteilungsmöglichkeiten darstellt. Mittels einer Poisson-Verteilung mit dem Mittelwert 6,25 ist zwar der wahrscheinlichste Ausgang der Wert 6 mit einer Wahrscheinlichkeit von knapp 16%, aber selbst der Wert 0 ist möglich; der Wert 5 mit einer Wahrscheinlichkeit von 15% fast genauso wahrscheinlich wie der Wert 6. Man spricht von statistischem Rauschen (noise), das zwar immer vorhanden ist, aber am deutlichsten im Tail zum Vorschein kommt. Die übliche Power-Law-Verteilung liefert große Werte für kleine k (wenige Autoren k mit vielen Papieren $f(k)$), wobei das Rauschen kaum wahrnehmbar ist. Für große k variiert der Wert von $f(k)$ zunehmend und wird sichtbar. Die Schwankungen sind als Zittern im Tail zu erkennen (s. Abb. 2). Man sollte dabei nicht den Fat-Tail mit Zittern verwechseln. Beide sind in Abb. 2 zu sehen. Der Fat-Tail ist eine normale Eigenschaft jeder Power-Law-Verteilung, da es wesentlich mehr Werte zum Ende der Verteilung gibt, die in Plot übereinander abgebildet werden und dadurch fetter wirken. Es handelt sich dabei um ein reines Darstellungsproblem.

Eine Möglichkeit, dem Zittern im Tail entgegenzuwirken, ist das sogenannte Binning, bei dem die Werte zu Gruppen zusammengefasst werden und so eine Darstellung mit geringeren Fehlern ermöglichen. Eine solche Fehlerbereinigung in der Darstellung und die Ermittlung der jeweiligen Variante bzw. die weiteren Feinheiten in der Bestimmung einer reinen Power-Law-Verteilung liegen außerhalb des Schwerpunktes dieser Arbeit. Eine Vielzahl an Methoden zur Bestimmung beschreiben z.B. Clauset et al. (2009). Es ist zudem gängige Praxis bei typischerweise unsauberen Verteilungen im IPP und in der Informationswissenschaft von Power-Laws zu sprechen, auch wenn diese bei genauerer Betrachtung nicht als reine Power-Laws angesehen werden dürften, sondern eine der obigen Varianten bzw. einer vollkommen anderen Verteilung zugerechnet werden müssten. Einer der Gründe hierfür ist die Einfachheit des Modells, das im Prinzip durch einen einzigen Wert beschrieben werden kann, den Exponenten. Der folgende Abschnitt beschäftigt sich mit den Verfahren zur Ermittlung des Exponenten.

5.9 Theoretische Ermittlung des Power-Law-Exponenten

Erlaubt der Plot einer gemessenen Verteilung durch Approximation einer Linie in der doppelt logarithmierten Darstellung, eine erste Einschätzung als Power-Law-Verteilung, kann der Exponent α ermittelt werden. Die einfachste Variante, den Exponenten zu bestimmen, ist der Einsatz des Standardverfahrens zur Ausgleichsrechnung, der Methode der kleinsten Quadrate. Während dieses Standardverfahren bei künstlich erzeugten Verteilungen ohne das statistische

Zittern im Tail zum richtigen Ergebnis führt, ist es nicht geeignet für realistische Daten, wobei hier auch das im vorherigen Abschnitt vorgestellte Binning helfen kann, den Fehler zu minimieren.

Newman (2005) stellt eine einfache Methode vor, um aus den Daten den Power-Law-Exponenten α ermitteln zu können, ohne dabei auf Hilfsverfahren wie Binning zurückgreifen zu müssen:

$$\alpha = 1 + n\left[\sum_{i=1}^{n} \ln\frac{x_i}{x_{min}}\right]^{-1}.$$

Die Werte für $x_i, i = 1\ldots n$ sind die gemessenen Werte von x, wobei x_{min} den kleinsten x-Wert bezeichnet. Allerdings kann das Verfahren nach Newman nur auf die kontinuierliche Variante angewendet werden und scheitert ebenfalls bei diskreten Werten. Clauset et al. (2009) und Egghe (2005, S. 387 ff.) stellen insgesamt drei Verfahren vor, um den Exponenten auch auf der Grundlage von diskreten Werten ermitteln zu können. Dort sind auch die Herleitungen zu den folgenden Formeln zu finden.

Die einfachste und schnellste Methode ist die sogenannte *Quick-and-Dirty*-Methode. Bei ihr wird $n_{max} = \infty$ definiert, A beschreibt die Gesamtzahl der Items und T die Gesamtzahl der Quellen (wobei $A > T$). Die beiden fehlenden Parameter zur Lösung sind somit der Exponent α und die Konstante K, die wie folgt ermittelt werden können:

$$\alpha = \frac{\ln\left(\frac{f(1)}{f(2)}\right)}{\ln 2}$$

sowie

$$K = f(1).$$

Die Annahme, dass K gleich der Anzahl von Quellen für ein Item ist, erleichtert die Bestimmung. Er kann direkt aus den vorhandenen Daten ausgelesen werden. Gleiches gilt für den Exponenten α. Allerdings ist kritisch anzumerken, dass bei dieser Quick-and-Dirty-Methode Messfehler in den ersten beiden Werten immense Auswirkungen haben, wobei die ersten beiden Werte meist wesentlich stabiler sind als dies z.B. im Tail der Fall ist.

Zwei zuverlässigere Methoden sind das Linear-Least-Square-Verfahren, das bereits von Lotka selbst angewendet wurde (Lotka, 1926) und eine Maximum-Likelihood Estimation, die wie folgt berechnet wird:

$$\frac{\sum_{n=1}^{n_{max}} f(n)\ln n}{\sum_{n=1}^{n_{max}} f(n)} = -\frac{\xi'(\alpha)}{\xi(\alpha)}.$$

Mittels $(-\xi'(\alpha))/(\xi(\alpha))$ kann der Exponent α errechnet werden. Bei ξ handelt es sich um die Riemann-Zèta-Funktion. Durch Einsetzen aller n diskreten Datenpunkte erhält man einen Wert, der in einer vorberechneten Wertetabelle nachgeschlagen werden kann. Beide Verfahren beruhen auf solchen Werteta-

bellen zur Bestimmung des Exponenten. Diese in dieser Arbeit wiederzugeben, erscheint nicht zielführend.

5.10 Ermittlung der Power-Law-Parameter in empirischen Daten

Zur Berechnung des Power-Law-Exponenten aus empirischen Daten (wie sie bei bibliometrischen Daten aus digitalen Bibliotheken vorliegen) gilt, dass x ein Vektor mit allen Beobachtungswerten ist. Die R-Funktion PLFIT.R ermittelt nun den Power-Law-Exponenten $\hat{\alpha}$, sodass $p(x) \sim x^{-\hat{\alpha}}$ für $x \geq \hat{x}_{min}$ gilt.

Das Programmbeispiel in Listing 1 generiert die Abb. 3 und zeigt den Aufruf von PLFIT zur Berechnung der Werte $\hat{\alpha}$, x_{min} und D. Die Berechnung ist laut Clauset et al. (2009) in zwei Schritte aufgeteilt:

1) Für jeden möglichen Wert von $\hat{x}_{min}$ wird ein $\hat{\alpha}$ geschätzt. Hierzu wird die Methode des maximum likelihood verwendet und anschließend die Kolmogorov-Smirnov goodness-of-fit-Statistik D berechnet.
2) Danach wird aus den möglichen $\hat{x}_{min}$ Werten derjenige ausgewählt, der den kleinsten D-Wert aller vorhandenen $\hat{x}_{min}$ besitzt.

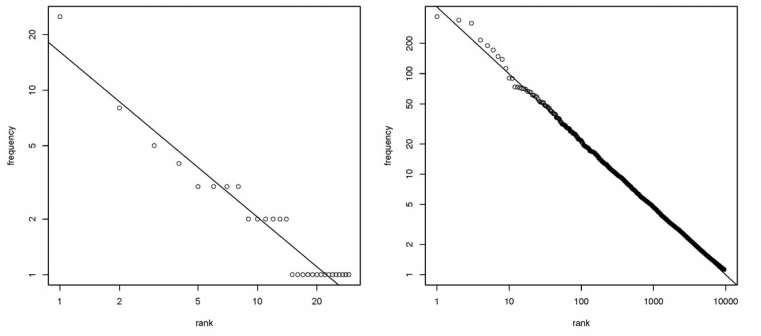
Die Autoren weisen darauf hin, dass dieses Verfahren keine Aussage über die Güte der Parameter oder die Validität der Schätzung macht. Dies ist allerdings systembedingt, da keine Power-Law-Verteilung, sondern eine z.B. Log-Normal-Verteilung vorliegen könnte.

Neben der Ermittlung des Exponenten und der $\hat{x}_{min}$ Parameter kann ebenfalls mittels des Kolomogorov-Smirnov-Tests die Güte der Schätzung (gof, *goodness-of-fit*) der Parameter errechnet werden. Bei der Berechnung des Tests werden sowohl die Werte gof als auch ein Wahrscheinlichkeitswert p ermittelt. Wenn p sehr groß ist (also nahe an 1), kann davon ausgegangen werden, dass ein Unterschied zwischen den empirischen Daten und dem errechneten Modell einer statistischen Schwankung (*statistical fluctuations*) zuzuschreiben ist. Sollte p sehr klein sein (nahe an 0), so kann davon ausgegangen werden, dass das Power-Law-Modell nicht zu den Daten passt. Dies ist zunächst kontraintuitiv, da allgemein bei der Überprüfung einer Null-Hypothese kleine Werte von p als aussagekräftig angesehen werden. In diesem Fall wird allerdings eine Hypothese überprüft, weshalb hohe Werte von p gewünscht werden. Clauset et al. (2009, 17) schlagen als Schwellenwert $p \geq 0{,}1$ vor. Liegt der ermittelte Wert p über dieser Schranke, so kann von einer Power-Law-Verteilung für die ermittelten Parameter ausgegangen werden.

Listing 1: Beispiel R-Skript zur Bestimmung der Power-Law-Exponenten und zum Plotten einer einfachen Verteilung

```
# Some frequencies (x)
freq1 <- c(25,8,5,4,3,3,3,3,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
freq2 <- sort((1-runif(1000))^(-1/(2.5-1)),decreasing=TRUE) # random scatter
# The rank numbers
index1 <- 1:length(freq1)
index2 <- 1:length(freq2)
# Plot frequencies on a double-logarithmic scale and draw an rough
# approximation (just for demonstration purposes)
plot(index1,freq1,log="yx",ylab="frequency",xlab="rank")
abline(lm(log10(freq1)~log10(index1)))
plot(index2,freq2,log="yx",ylab="frequency",xlab="rank")
abline(lm(log10(freq2)~log10(index2)))
# Now that we know what we are dealing with, compute the exponent
plfit(freq1)
[1] "(plfit) Warning : finite-size bias may be present"
$xmin
[1] 2
$alpha
[1] 2.35
$D
[1] 0.05350649
```

Abb. 3: Zwei Plots für Power-Law-Verteilungen und deren Approximation mit R



6. Diskussion

Zunächst ist festzuhalten, dass die in diesem Beitrag ermittelten Werte der Produktivitätsanalyse nicht mit Relevanz gleichzusetzen sind. Relevanz ist als ein multidimensionales, dynamisches und komplexes Konzept zu verstehen. Die ermittelten Produktivitätszahlen innerhalb eines IPP beziffern ausschließlich die Anzahl der von einer Informationsquelle (Sources, Zeitschriften, Auto-

ren etc.) produzierten Informationseinheiten (Items, welche von den Sources produziert werden, z.B. Artikel). Gleiches gilt für andere informetrische Analysen, z.B. mittels Kookkurrenzanalyse: Mit ihrer Hilfe kann streng genommen nur die Häufigkeit eines gemeinsamen Auftretens zweier Informationseinheiten (Worte in einem Text, Autorennamen, etc.) innerhalb eines Dokuments gemessen werden. Zwar wird die Annahme zugrunde gelegt, dass eine strukturelle bzw. semantische Abhängigkeit zwischen den Informationseinheiten besteht, doch ist dies nicht immer zwangsläufig korrekt. Die Verfahren sind per se nicht in der Lage, relevante Dokumente zu liefern.

Allerdings ermöglichen die hier zusammengetragenen Methoden Verbindungen zwischen beiden Disziplinen herzustellen. So lassen sich auf Grundlage der informetrischen Analysen sogenannte Mehrwertdienste entwickeln, die die Probleme von Benutzern digitaler Bibliotheken zu lösen versuchen. Die Analyse der Power-Law-Verteilungen kann beispielsweise dazu verwendet werden ein alternatives Ranking anzubieten, dass über einfache Term-Gewichtungen hinausgeht. Weiterhin können Recommendersysteme gebaut werden, die den Nutzer bei der Formulierung seines Informationsbedürfnisses aktiv unterstützen können. Eine umfangreise Evaluation solcher Systeme und deren Vor- und Nachteile sind bei Schaer (2013b) in den Teilen II und III nachzulesen.

Es stellt sich natürlich die Frage, ob sich die informetrische Analyse allgemein und die Produktivitätsanalyse im Speziellen – abseits von der eigentlichen Machbarkeit – überhaupt konzeptionell eignen, um das Information Retrieval anzureichern. Eine ähnliche Diskussion ist in der Zitationsanalyse zu beobachten, der anderen großen informetrischen Disziplin, die in diesem HSR Focus bewusst ausgeklammert wurde. Wie von Ingwersen (2012) beschrieben, ist auch die Anzahl an Zitationen kein Relevanzmerkmal: „Citations do not signify relevance! But the number of citations signify utility in a particular work/context" (Ingwersen, 2012, Minute 8).

Eine hohe Anzahl an Zitationen ist kein Hinweis auf (topical) Relevanz, aber auf die Nützlichkeit in einer jeweiligen Situation. So kann z.B. durch das Zitieren einer Arbeit auch ein negatives Beispiel angeführt werden („... wie von Quelle X gezeigt ist der gewählte Ansatz A falsch, daher wird der alternative Ansatz B verfolgt ..."). In diesem Fall wäre das gewählte Zitat ein Hinweis auf eine offensichtlich falsche Quelle, die objektiv betrachtet keine relevanten Informationen enthält. Mit Hilfe eines Zitationsindex kann laut Ingwersen folglich nur etwas über die „social utility" oder die „academic (re)cognition" ausgesagt werden. Er grenzt dies klar von Relevanz ab.

Diese konzeptionellen Schwierigkeiten zeigen sich auch in Arbeiten von Larsen (2004), der Referenzen und Zitationen für den Retrievalprozess nutzbar machen wollte und keine Verbesserungen im Vergleich zu einem regulären Best-Match-Verfahren feststellen konnte. Larsen sieht einen konkreten Schwachpunkt:

> References and citations are fundamentally different from conventional term-based representations, e.g., they retrieve different documents for the same request, and the overlap between the two is typically small (Larsen 2004, 234).

Zum Missverhältnis von Zitationen und Relevanz kommt der Aspekt der Nicht-Zitation („uncitedness“) vieler Werke, die den Einsatz von Verfahren der Zitationsanalyse im IR erschweren. Diese Probleme bestehen bei der Produktivitätsanalyse nicht, da hier mit absoluten Häufigkeiten von vorhandenen Daten gearbeitet wird, die nicht von der direkten Verknüpfung mit anderen Daten abhängig sind. Der Vergleich mit der Zitationsanalyse ist an dieser Stelle legitim, da diese mit den gewählten informetrischen Verfahren einige grundlegende Schwächen gemein hat – genannt seien hier die starke Abhängigkeit von der betrachteten Disziplin und der Qualität bzw. fachlichen Abdeckung der Datengrundlage. Diese Probleme sind allgegenwärtig bei informetrischen Analysen.

Durch die gezielte Analyse der Häufigkeitsverteilungen für kookkurrierende Dokumentattribute, deren Untersuchung hinsichtlich des Vorliegens einer Power-Law-Verteilung und deren Anwendung als Filter für die entwickelten Systeme in Schaer (2013b), konnte die Retrievalleistung der betrachteten Systeme hingegen die Baseline schlagen oder zumindest an sie heranreichen. Dies war sowohl für die Dokumentattribute Autorennamen, Zeitschriftenzugehörigkeit (ISSN-Codes) und kontrollierte Verschlagwortung mit Thesaurustermen für als auch für die Publikationsquelle zu beobachten. Gleichzeitig konnte die Analyse der τ-Werte zeigen, dass das alternative Ranking für diese Systeme eine substantiell veränderte Ergebnisliste generiert und somit dem Benutzer einen neuen Blick auf den Datenbestand liefern konnte.

Dies ist zunächst erstaunlich, da die gewählten Verfahren und die eingesetzten Filter- und Gütekriterien keinen direkten Bezug zur Relevanz haben. Die Interpretation der Power-Law-Exponenten als Filterkriterium für das Retrieval ist in dieser Anwendung ungewöhnlich, da dieser in der Literatur nicht für praktische Schlussfolgerungen genutzt wird. Der Grad des Gefälles (slope) der Power-Law-Funktion wird allgemein als die Stärke der Unterschiede in einer Verteilung interpretiert. Eine akzeptierte Interpretation ist das Gefälle z.B. als ein Gradmesser der „Fairness“ in einem Netzwerk. Eine exakte Quantifizierung der Power-Law-Verteilungen hat aber meist nur einen beschreibenden Charakter oder wird allgemein als nicht notwendig eingestuft.

Aber auch die Interpretation in die andere Richtung ist möglich, die nicht den Mehrwert der Informetrie für das IR betrachtet, sondern den des IR für die Informetrie. Die empirische Überprüfung von informetrischen Modellen stellt ein Problem dar, dem mit dem rigiden Evaluationsframework des Cranfield-Paradigmas begegnet werden soll:

> The science models studied are therefore verified as expressive models of science, as an evaluation of retrieval quality is seen as a litmus test of the adequacy of the models investigated (Mutschke u.a. 2011, 362).

Dieser Lackmustest (litmus test) zielt darauf ab, die Vorteile der IR-Laborevaluation für die Informetrie zu nutzen. Wie auch in der Chemie wird hier der Begriff des Lackmustest nur für eine grobe Abschätzung verwendet (z.B. ob eine saure oder basische Lösung vorliegt, bzw. ob die Gültigkeit eines informetrischen Modells plausibel erscheint oder nicht). Die Gültigkeit der Modelle soll durch die Anwendung im IR (zumindest teilweise) belegt werden. Ein auf diesen Modellen basiertes IR-System kann bei erfolgreicher Evaluation mit einer IR-Testkollektion als ein Indikator für die Gültigkeit bzw. Plausibilität des betrachteten Modells gesehen werden. Die Verknüpfung von IR-Relevanzurteilen und den Ergebnissen einer informetrischen Analyse ist allerdings nur ein Indikator, der genutzt werden könnte, um einen Hinweis auf Korrektheit eines zugrundeliegenden Modells zu erlangen kann.

In diesem Papier wurden die grundlegenden methodischen Ansätze des Information Retrieval und der Informetrie zusammengefasst, die genutzt werden können, um neue Retrievalmöglichkeiten für den Anwendungsbereich der digitalen Bibliotheken zu entwerfen und zu evaluieren, wie dies zum Beispiel in Schaer (2013b) ausführlich geschehen ist. Es zeigt sich, dass durch die Anwendung von informetrischen Analysen ein tatsächlicher Mehrwert für den Endanwender entstehen kann, der sich darüber hinaus auch mit den hier vorgestellten IR-Kennzahlen messen und quantifizieren lässt.

Dass es über die reine Machbarkeit hinaus ein starkes Interesse an der Verquickung der beiden Disziplinen gibt, zeigen unter anderem aktuelle Workshops wie „Computational Scientometrics“ auf der CIKM 2013 (eine einflussreiche IR-Konferenz) und „Combining Bibliometrics and Information Retrieval“ (bei ISSI 2013, die größte Informetrie-Konferenz weltweit). Es ist daher zu hoffen, dass die vielen potentiellen Überschneidungen der Disziplinen in der nahen Zukunft zu einer Mehrzahl an tatsächlichen Mehrwertdiensten für Nutzer führen, so dass die eingangs beschriebenen Suchprobleme in digitalen Bibliotheken etwas von ihrer Schärfe verlieren.

References

Adamic, Lada A. 2000. *Zipf, Power-laws, and Pareto – a ranking tutorial.* <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html> (Zugegriffen: 18. März 2011).

Alonso, Omar, Ralf Schenkel, und Martin Theobald. 2010. Crowdsourcing Assessments for XML Ranked Retrieval. In *Advances in Information Retrieval*, hg. v. Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, und Keith van Rijsbergen, 5993: 602-6. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. (Zugegriffen: 4. Mai 2012).

Anderson, Rick. 2011. The Crisis in Research Librarianship. *The Journal of Academic Librarianship* 37 (4): 289-90 <doi:10.1016/j.acalib.2011.04.001>.

Arms, William. 2000. *Digital libraries*. Cambridge, MA: MIT Press <http://www.cs.cornell.edu/wya/diglib/>.

Bao, Shenghua, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, und Zhong Su. 2007. Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web*, 501-10. WWW '07. New York, NY: ACM <doi:10.1145/1242572.1242640>, <http://portal.acm.org/citation.cfm?doid=1242572.1242640>.

Beel, Jöran, und Bela Gipp. 2009. Google Scholar's Ranking Algorithm: An Introductory Overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, 230-41. International Society for Scientometrics and Informetrics, Juli.

Benz, Dominik, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, und Gerd Stumme. 2010. The social bookmark and publication management system bibsonomy. *The VLDB Journal* 19 (6): 849-75 <doi:10.1007/s00778-010-0208-4>.

Te Boekhorst, Peter, Matthias Kayß, und Roswitha Poll. 2003. *Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung: Teil I: Informationsverhalten und Informationsbedarf der Wissenschaft*. Universitäts- und Landesbibliothek Münster, infas Institut für angewandte Sozialwissenschaft GmbH <http://www.dfg.de/download/pdf/foerderung/programme/lis/ssg_bericht_teil_1.pdf> (Zugegriffen: 10. Dezember 2012).

Borlund, Pia. 2003. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology* 54 (10): 913-25 <doi:10.1002/asi.10286>.

Bradford, Samuel C. 1934. Sources of information on specific subjects. *Engineering* 137: 85-6.

Buckley, Chris. 2009. Why current IR engines fail. *Information Retrieval* 12 (6): 652-65. <doi:10.1007/s10791-009-9103-2>.

Buckley, Chris, und Ellen M Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 33-40. SIGIR '00. New York, NY: ACM. <doi:10.1145/345508.345543>.

Buckley, Chris und Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 25-32. SIGIR '04. New York, NY: ACM <doi:10.1145/1008992.1009000>, <http://doi.acm.org/10.1145/1008992.1009000>.

Calhoun, Karen, Joanne Cantrell, Peggy Gallagher, und Janet Hawk. 2009. *Online Catalogs: What Users and Librarians Want*. OCLC Report. OCLC Report. Dublin, Ohio: OCLC <http://www.oclc.org/reports/onlinecatalogs/default.htm> (Zugegriffen: 23. Juni 2011).

Chowdhury, Gobinda G. 2010. *Introduction to modern information retrieval*, 3rd ed. London: Facet.

Clauset, Aaron, Cosma Rohilla Shalizi, und M. E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *Society for Industrial and Applied Mathematics Review* 51 (4): 661-703 <doi:10.1137/070710111>.

Cleverdon, Cyril W. 1960. The ASLIB Cranfield Research Project on the Comparative Efficiency of Indexing Systems. *Aslib Proceedings* 12 (12): 421-31 <doi: 10.1108/eb049778>.

Crook, Thomas, Brian Frasca, Ron Kohavi, und Roger Longbotham. 2009. Seven pitfalls to avoid when running controlled experiments on the web. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '09: 1105-14 <doi:10.1145/1557019.1557139>.

Csányi, Gábor, und Balázs Szendrodblaci. 2004. Structure of a large social network. *Physical Review E* 69 (3): 036131 <doi:10.1103/PhysRevE.69.036131>.

Egghe, Leo. 2005. *Power Laws in the Information Production Process: Lotkaian Infometrics*. Library and Information Science Series. Oxford: Elsevier.

Egghe, Leo. 2009. Lotkaian informetrics and applications to social networks. *Bulletin of the Belgian Mathematical Society – Simon Stevin* 16 (4): 689-703.

Fain, Daniel C., und Jan O. Pedersen. 2006. Sponsored search: A brief history. *Bulletin of the American Society for Information Science and Technology* 32 (2): 12-3 <doi:10.1002/bult.1720320206>.

Fang, Hui, Tao Tao, und Chengxiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Transactions on Information Systems* 29 (2): 7:1-7:42 <doi:10.1145/1961209.1961210>.

Ferber, Reginald. 2003. *Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. Heidelberg: dpunkt Verlag.

Fox, Edward A., Marcos Andre Goncalves, und Rao Shen. 2012. *Theoretical Foundations for Digital Libraries: The 5S Approach*. Morgan & Claypool Publishers.

Fuhr, Norbert. 2004. Theorie des Information Retrieval I: Modelle. In *Grundlagen der praktischen Information und Dokumentation*, 5. Aufl., hg. v. Rainer Kuhlen, T. Seeger, und D. Strauch, 207-14. München: Saur.

Glänzel, Wolfgang. 2003. *Bibliometrics as a Research Field* <http://nsdl. niscair.res.in/bitstream/123456789/968/1/Bib_Module_KUL.pdf> (Zugegriffen: 20. Dezember 2011).

Glänzel, Wolfgang. 2007. Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics* 1 (1): 92-102 <doi:10.1016/j.joi.2006.10.001>.

Glänzel, Wolfgang, Frizo Janssens, und Bart Thijs. 2008. A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics* 79 (1): 109-29 <doi:10.1007/s11192-009-0407-1>.

Google Scholar. 2011. *About Google Scholar* <http://scholar.google.com/scholar/about.html?hl=en> (Zugegriffen: 4. August 2011).

Hardy, Michael. 2010. Pareto's Law. *The Mathematical Intelligencer* 32 (3): 38-43 <doi:10.1007/s00283-010-9159-2>.

Harzing, Anne-Wil. 2010. *The Publish or perish book : your guide to effective and responsible citation analysis*. Melbourne: Tarma Software Research Pty Ltd.

Havemann, Frank. 2009. *Einführung in die Bibliometrie*. Berlin: Gesellschaft für Wissenschaftsforschung <http://edoc.hu-berlin.de/oa/books/reMKADKkid1Wk/PDF/20uf7RZtM6ZJk.pdf>.

He, Zi-Lin. 2009. International collaboration does not have greater epistemic authority. *Journal of the American Society for Information Science and Technology* 60 (10): 2151-64 <doi:10.1002/asi.v60:10>.

Hienert, Daniel, Philipp Schaer, Johann Schaible, und Philipp Mayr. 2011. A Novel Combined Term Suggestion Service for Domain-Specific Digital Libraries. In *Research and Advanced Technology for Digital Libraries*, hg. v. Stefan Gradmann, Francesca Borri, Carlo Meghini, und Heiko Schuldt, 6966: 192-203. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer (Zugegriffen: 13. September 2011).

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102 (46): 16569-72 <doi:10.1073/pnas.0507655102>.

Hjørland, Birger. 2000. Library and information science: practice, theory, and philosophical basis. *Information Processing and Management* 36 (3): 501-31 <doi:10.1016/S0306-4573(99)00038-2>.

Hotho, Andreas, Robert Jäschke, Christoph Schmitz, und Gerd Stumme. 2006. Information Retrieval in Folksonomies: Search and Ranking. In *The Semantic Web: Research and Applications*, hg. v. York Sure und John Domingue, 4011: 411-26. Lecture Notes in Artificial Intelligence. Heidelberg: Springer, Juni <doi:10.1007/11762256_31>, <http://.kde.cs.uni-kassel.de/hotho>.

Ingwersen, Peter. 2012. *Bibliometrics/Scientometrics and IR A methodological bridge through visualization*. Vortrag gehalten auf der PROMISE Winter School 2012, Januar, Zinal, Valais <http://www.promise-noe.eu/documents/10156/028a48d8-4ba8-463c-acbc-db75db67ea4d> (Zugegriffen: 12. August 2012).

Ingwersen, Peter, und Kalervo Järvelin. 2005. *The turn – integration of information seeking and retrieval in context*. Dordrecht: Springer.

Jacsó, Péter. 2008. Google Scholar revisited. *Online Information Review* 32 (1): 102-14 <doi:10.1108/14684520810866010>.

Järvelin, Kalervo, und Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20 (4): 422-46 <doi:10.1145/582415.582418>.

Jordy, Matthew L., Eileen L. McGrath, und John B. Rutledge. 1999. Book Reviews As a Tool for Assessing Publisher Reputation. *College & Research Libraries* 60 (2): 132-42. (Zugegriffen: 3. Mai 2011).

Juran, Joseph M. 1954. Universals in management planning and controlling. *Management Review* 43 (11): 748-61.

Kahn, Robert E., und Vinton G. Cerf. 1988. *The Digital Library Project Volume I: The World of Knowbots (DRAFT): An Open Architecture For a Digital Library System and a Plan For Its Development*. Corporation for National Research Initiatives <http://hdl.handle.net/4263537/2091>.

Kelly, Diane. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3 (1): 1-224 <doi:10.1561/1500000012>.

Kleinberg, Jon M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46 (5): 604-32 <doi:10.1145/324133.324140>.

Kürsten, Jens, und Maximilian Eibl. 2011. A Large-Scale System Evaluation on Component-Level. In *Advances in Information Retrieval*, hg. v. Paul Clough, Colum Foley, Cathal Gurrin, Gareth Jones, Wessel Kraaij, Hyowon Lee, und Vanessa Mudoch, 6611: 679-82. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. (Zugegriffen: 12. Juni 2012).

Larsen, Birger. 2004. *References and citations in automatic indexing and retrieval systems – experiments with the boomerang effect*. PhD Thesis, Copenhagen, Denmark: Department of Information Studies, Royal School of Library and Information Science <http://pure.iva.dk/files/31034810/birger_larsen_phd.pdf>.

Lewandowski, Dirk. 2009. Ranking library materials. *Library Hi Tech* 27 (4): 584-93 <doi:10.1108/07378830911007682>.

Lotka, Alfred J. 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 16 (12): 317-23.

Mandl, Thomas. 2006. *Die automatische Bewertung der Qualität von Internet-Seiten im Information Retrieval*. Habitilationsschrift, Universität Hildesheim.

Manning, Christopher D., Prabhakar Raghavan, und Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>.

Metzler, Donald. 2011. *A Feature-Centric View of Information Retrieval*. Bd. 27. The Information Retrieval Series. Berlin, Heidelberg: Springer.

Milojević, Staša. 2010. Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology* 61: 2417-25 <doi:10.1002/asi.v61:12>.

Morris, Steven. A., und Gary G. Yen. 2004. Crossmaps: Visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences* 101 (1): 5291-6 <doi:10.1073/pnas.0307604100>.

Mutschke, Peter, Philipp Mayr, Philipp Schaer, und York Sure. 2011. Science models as value-added services for scholarly information systems. *Scientometrics* 89 (1): 349-64 <doi:10.1007/s11192-011-0430-x>.

Neal, Diane Rasmussen, Hg. 2012. *Indexing and Retrieval of Non-Text Information*. Knowledge and Information / Studies in Information Science. De Gruyter – Saur.

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46: 323-51 <doi:10.1080/00107510500052444>.

Page, Lawrence, Sergey Brin, Rajeev Motwani, und Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab <http://ilpubs.stanford.edu:8090/422/>.

Peters, Isabella. 2010. *Folksonomies. Indexing and Retrieval in Web 2.0*. Berlin: De Gruyter – Saur.

Peters, Isabella, und Wolgang G. Stock. 2010. Power tags in information retrieval. *Library Hi Tech* 28 (1): 81-93.

Potter, William Gray. 1981. Lotka's Law Revisited. *Library Trends* 30 (1): 21-39. (Zugegriffen: 12. April 2011).

Redner, Sidney. 2010. On the meaning of the h-index. *Journal of Statistical Mechanics: Theory and Experiment* 2010 (16. März): L03005 <doi:10.1088/1742-5468/2010/03/L03005>.

Van Rijsbergen, C. J. 1974. Foundation of Evaluation. *Journal of Documentation* 30 (4): 365-73 <doi:10.1108/eb026584>.

Robertson, Stephen E. 1977. Theories and Models in Information Retrieval. *Journal of Documentation* 33 (2): 126-48.

Robertson, Stephen E., Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, und Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of 3rd Text Retrieval Conference*, 109-2. <http://dblp.uni-trier.de/rec/bibtex/conf/trec/RobertsonWJHG94> (Zugegriffen: 10. Oktober 2012).

Robertson, Stephen E., und Hugo Zaragoza. 2010. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3 (4): 333-89 <doi:10.1561/1500000019>.

Sanderson, Mark. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4 (4): 247-375 <doi:10.1561/1500000009>.

Sanderson, Mark, und Ian Soboroff. 2007. Problems with Kendall's tau. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 839-40. SIGIR '07. New York, NY: ACM <doi:10.1145/1277741.1277935>, <http://doi.acm.org/10.1145/1277741.1277935>.

Saracevic, Tefko. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology* 58 (13): 1915-33 <doi:10.1002/asi.v58:13>.

Saracevic, Tefko. 2009. Introduction: the framework for digital library evaluation. In *Evaluation of Digital Libraries: An Insight to Useful Applications and Methods*, hg. v. Giannis Tsakonas und Christos Papatheodorou, 1-13. Chandos Publishing (Oxford) Ltd.

Schaer, Philipp. 2013a. Applied Informetrics for Digital Libraries: An Overview on Problems, Foundations and Current Approaches. *Historical Social Research* 38 (3): 267-81.

Schaer, Philipp. 2013b. *Der Nutzen informetrischer Analysen und nicht-textueller Dokumentattribute für das Information Retrieval in digitalen Bibliotheken*. Dissertation, Koblenz: Universität Koblenz-Landau, 27. Mai <http://kola.opus.hbz-nrw.de/frontdoor.php?source_opus=896&la=de> (Zugegriffen: 19. August 2013).

Schaer, Philipp, Philipp Mayr, und Thomas Lüke. 2012. Extending Term Suggestion with Author Names. In *Theory and Practice of Digital Libraries*, hg. v. Panayiotis Zaphiris, George Buchanan, Edie Rasmussen und Fernando Loizides, 7489: 317-22. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer.

Scharnhorst, Andrea, und Eugene Garfield. 2010. Tracing scientific influence. *Journal of Dynamics of Socio-Economic Systems* 2 (1): 1-33.

Schlögl, Christian, und Juan Gorraiz. 2012. Sind Downloads die besseren Zeitschriftennutzungsdaten? Ein Vergleich von Download und Zitationsindikatoren. *Zeitschrift für Bibliothekswesen und Bibliographie* 59 (2): 87-95.

Seadle, Michael. 2009. Digitale Bibliothek. In *Lexikon der Bibliotheks- und Informationswissenschaft (LBI)*, hg. v. Konrad Umlauf und Stefan Gradmann, 3: 216-7. Stuttgart: Anton Hiersemann Verlag.

Siegfried, Doreen, und Elisabeth Flieger. 2011. *World Wide Wissenschaft – Wie professionell Forschende im Internet arbeiten*. ZBW – Leibniz-Informationszentrum Wirtschaft <http://www.zbw.eu/presse/pressemitteilungen/docs/world_wide_wissenschaft_zbw_studie.pdf>.

De Solla Price, Derek J. 1963. *Little Science, Big Science*. New York: Columbia University Press.

Song, Ruihua, Qingwei Guo, Ruochi Zhang, Guomao Xin, Ji-Rong Wen, Yong Yu, und Hsiao-Wuen Hon. 2011. Select-the-Best-Ones: A new way to judge relative relevance. *Information Processing & Management* 47 (1): 37-52 <doi:10.1016/j.ipm.2010.02.005>.

Spärck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1): 11-21.

Sühl-Strohmenger, Wilfried. 2008. *Digitale Welt und Wissenschaftliche Bibliothek-Informationspraxis im Wandel: Determinanten, Ressourcen, Dienste, Kompetenzen: eine Einführung*. Wiesbaden: Harrassowitz.

Tague-Sutcliffe, Jean. 1992. An introduction to informetrics. *Information Processing & Management* 28 (1): 1-3 <doi:10.1016/0306-4573(92)90087-G>.

Umstätter, Walther. 1999. *Zum Thema Lotka's law*. 23. Februar <http://www.ib.hu-berlin.de/~wumsta/price52.html>.

Umstätter, Walther. 2003. *Das Principle of Least Effort in der Wissenschaft*. Vortrag im Berliner Bibliothekswissenschaftlichen Kolloquium gehalten auf der Berliner Bibliothekswissenschaftlichen Kolloquiums (BBK), 7. Oktober, Saur-Bibliothek des Instituts für Bibliothekswissenschaft <http://www.ib.hu-berlin.de/~wumsta/infopub/lectures/leasteffort03a.pdf>.

Voorhees, Ellen M. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, 315-23. New York, NY: ACM <doi:10.1145/290941.291017>.

Voorhees, Ellen M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management* 36 (5): 697-716 <doi:10.1016/S0306-4573(00)00010-8>.

Voß, Jakob. 2007. Visualisierung der ZDB. *Verbund-Wiki GBV*. 3. April <http://www.gbv.de/wikis/cls/Visualisierung_der_ZDB> (Zugegriffen: 14. Dezember 2011).

Webber, William, Alistair Moffat, Justin Zobel, und Tetsuya Sakai. 2008. Precision-at-ten considered redundant. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 695-6. SIGIR '08. Singapore: ACM <doi:10.1145/1390334.1390456> .

White, Howard D. 2007. Combining bibliometrics, information retrieval, and relevance theory, Part 1: First examples of a synthesis: Research Articles. *Journal of the American Society for Information Science and Technology* 58: 536-59 <doi:10.1002/asi.v58:4>.

Wong, William, Hanna Stelmaszewska, Balbir Barn, Nazlin Bhimani, und Barn Sukhbinder. 2010. *JISC user behaviour observational study: User behaviour in resource discovery*. JISC <http://www.jisc.ac.uk/publications/programmerelated/2010/ubirdfinalreport.aspx> (Zugegriffen: 23. Juni 2011).

Yin, Li-chun, Hildrun Kretschmer, Robert A Hanneman, und Ze-yuan Liu. 2006. Connection and stratification in research collaboration: an analysis of the COLLNET network. *Information Processing & Management* 42 (6): 1599-613 <doi:10.1016/j.ipm.2006.03.021>.

Zanardi, Valentina, und Licia Capra. 2008. Social ranking: uncovering relevant content using tag-based recommender systems. In *Proceedings of the 2008 ACM conference on Recommender systems*, 51-8. RecSys '08. Lausanne, Switzerland: ACM <doi:10.1145/1454008.1454018>.

Zhang, Yan. 2008. Undergraduate students mental models of the Web as an information retrieval system. *Journal of the American Society for Information Science and Technology* 59 (13): 2087-98 <doi:10.1002/asi.v59:13>.

Zhu, Jianhan, Dawei Song, und Stefan Rüger. 2009. Integrating multiple windows and document features for expert finding. *Journal of the American Society for Information Science and Technology* 60 (4): 694-715 <doi:10.1002/asi.21012>.
Zipf, George K. 1949. *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.
Zobel, Justin, Alistair Moffat, und Laurence A. F Park. 2009. Against recall: is it persistence, cardinality, density, coverage, or totality? *SIGIR Forum* 43 (1): 3-8 <doi:10.1145/1670598.1670600>.