Alexander Mehler

Automatische Synthese Internet-basierter Links für digitale Bibliotheken

Abstract

Dieser Beitrag behandelt Verfahren zur automatischen Erzeugung von Hyperlinks, wie sie im WWW für die Informationssuche bereitstehen. Dabei steht die Frage im Vordergrund, auf welche Weise bestehende Verfahren suchrelevante Dokumente bestimmen und von diesen aus inhaltsverwandte Dokumente verlinken. Dieser Gegenstand verbindet den Bereich des klassischen Information Retrievals (IR) mit einem Anwendungsgebiet, das in der Wissenschaftskommunikation unter dem Stichwort der digitalen Bibliothek unter Nutzbarmachung des Hyperlink-basierten Browsings firmiert. Ein Beispiel hierfür bildet die digitale Bibliothek CiteSeer (Lawrence et al. 1999), welche das Boolesche Retrieval dadurch erweitert, dass ausgehend von Treffern einer Suche jene Dokumente per Link angesteuert werden können, welche die aufgefundenen Dokumente zitieren oder von diesen zitiert werden. CiteSeer ist also ein System, welches das Schlagwort-basierte Querying im Rahmen des klassischen IRs mit dem Hypertext-basierten Browsing von Zitaten verknüpft, und zwar zu dem Zweck, die Suche wissenschaftlicher Dokumente zu erleichtern. Darüber hinaus verwendet es die unter dem Stichwort des Vektorraummodells bekannt gewordene Technologie für den wortbasierten Vergleich von Texten. Der Beitrag setzt an dieser Stelle an. Er argumentiert, dass Verfahren bereitstehen, welche die Anforderung nach inhaltsorientiertem Retrieval mit dem inhaltsorientierten Browsing verbinden, mit der Forderung also, dass Hyperlinks, die E-Texte als digitalisierte Versionen von (wissenschaftlichen) Dokumenten verknüpfen (Storrer 2002), Inhalts- und nicht nur Zitat-basiert sind.

1. Einleitung

Die Beschäftigung mit der automatischen Erzeugung von Hyperlinks steht in Zusammenhang mit der Entwicklung digitaler Bibliotheken, welche

den inhaltsorientierten Zugriff auf digitalisierte Dokumentbestände über bestehende Suchmaschinentechnologien hinaus anstreben (Lossau 2004). Dabei stellt sich die nahe liegende Frage, ob eine automatische und zugleich inhaltsbasierte Verlinkung von Texten überhaupt möglich ist. Setzt dies nicht, so ist zu fragen, eine Interpretationsleistung voraus, die gar nicht automatisierbar ist? An dieser Stelle folgt der Beitrag die Auffassung, wonach gar keine andere Wahl besteht, als diese Aufgabe anzugehen. Hierfür sprechen unter anderem zwei Argumente:

- Kobayashi/Takeda 2000 beschreiben den exponentiellen Zuwachs der Zahl online verfügbarer Webpages, welche wiederum mit der Zahl der Suchanfragen an hochfrequentierte Suchdienste korreliert. Diesem Zuwachs stehen fehlerhaft annotierte, häufig geänderte Pages ebenso gegenüber wie ungültige und irrelevante Links. Nach Kobayashi/Takeda wird jede Seite im Schnitt alle 75 Tage und 40 % der Seiten jeden Monat, manche mehrmals am Tag geändert. Auch wenn vergleichbare Daten für den Bereich der Internet-basierten Wissenschaftskommunikation (noch) fehlen, ist von tendenziell ähnlichen Verhältnissen auszugehen. Es ist somit dringend erforderlich, den Zugang zu Dokumenten über den Booleschen Nachweis von Suchtermen hinaus zu erweitern, wobei die schiere Menge der Dokumente und die damit verbundene Update-Problematik ihre automatische Analyse unabdingbar macht.
- Aus anwendungsorientierter Sicht stellt sich die Frage nach der angemessenen Verarbeitung immer größerer Mengen digitalisierter Texte. Auch wenn Hypertexte heute nicht länger ernsthaft als Ersatz für „lineare Texte“ diskutiert werden, besteht nach wie vor das Problem einer effizienten computerbasierten Verarbeitung großer Textmengen. Dabei erweist sich das WWW mit seinem einfachen Hypertextmodell als ein erfolgreiches Medium der computerbasierten Informationsverarbeitung. Folgerichtig liegt es nahe, dieses Medium auch zur Exploration solcher Textmengen zu nutzen, die als zusammenhangslose E-Texte existieren oder überhaupt erst digitalisiert werden müssen.

Bei aller gebotenen Skepsis besteht kaum eine andere Wahl, als das Instrumentarium der automatischen Textanalyse, wie es innerhalb von Informatik, Computerlinguistik und Texttechnologie entwickelt wurde, zur Vorverarbeitung digitalisierter wissenschaftlicher Dokumente einzusetzen, um letztere Aufgabe der Massendatenanalyse zu bewältigen. Anstatt also diese Aufgabe den Ingenieurswissenschaften zu überantworten und sich von etwaigen Ergebnissen überraschen zu lassen, ist nach den Grenzen einer solchen automatischen Zeichenverarbeitung zu fragen,

wie sie erst im Rahmen einer zeichentheoretisch fundierten und zugleich technologisch orientierten Disziplin zu bestimmen sind. Die aktuelle Entwicklung im Bereich von Suchmaschinen, welche sich zunehmend der Verfügbarmachung wissenschaftlicher Dokumente widmen, wie das jüngste Beispiel eines der führenden Anbieters auf diesem Markt zeigt[1], unterstreicht die Wichtigkeit, die Kontrolle über das technisch Machbare und das hiervon Realisierte zu behalten. In diese Richtung zielt der vorliegende Beitrag, wobei die folgenden Abschnitte das bestehende Instrumentarium anhand ausgewählter Beispiele kritisch beleuchtet, und zwar stets unter Rekurs auf die Anforderung eines inhaltsorientierten, Hypertext-basierten Dokumentzugriffs, wie ihn digitale Bibliotheken anstreben. Dabei werden fast ausnahmslos die in der Informatik üblichen Termini ‚Dokument‘ und ‚(Index-)Term‘ anstelle von ‚Text‘ und ‚(lexikalischer) Komponente‘ verwendet. Ferner sei angemerkt, dass sich dieser Beitrag *nicht* mit der textlinguistischen Fundierung von Hypertexten beschäftigt. Diese Frage wird in Mehler 2005 aufgegriffen und unter Rekurs auf eine relationenalgebraische Rekonstruktion des Begriffs der intertextuellen Kohärenzrelation beantwortet.

2. Inhaltsbasierte Verlinkung

Inhaltsbasierte Verlinkung von Dokumenten bedeutet unter anderem, diese Dokumente auch dann miteinander zu verknüpfen, wenn sie zwar oberflächenstrukturell verschieden sind, indem sie etwa in verschiedenen Sprachen verfasst sind, dafür aber ähnliche Inhalte behandeln. Für Ähnlichkeitsrelationen von Texten stehen mit der Textfunktion, dem Textinhalt, der (logischen) Textstruktur und der äußeren Textgestalt vier Bezugsgrößen bereit (Brinker 1992). Aus Gründen der Einfachheit beschränkt sich dieser Artikel auf die Frage nach der Berechnung der inhaltlichen, thematischen Ähnlichkeit. Struktur- und Funktions-orientierte Gesichtspunkte solcher Ähnlichkeitsrelationen werden in Mehler 2002b und 2004a behandelt.

Insoweit Textähnlichkeitsrelationen mit Hilfe von Hyperlinks digitalisiert werden, ist die so genannte Explikationshypothese als theoretische Basis der Verlinkung heranzuziehen. Diese Hypothese besagt, dass Links dazu dienen können, die Ausdrucksseiten von Kohäsions- und Kohärenz-

1 Dabei handelt es sich um *Google Scholar*, ein System, das ausschließlich wissenschaftliche Literatur auffinden helfen können soll.

relationen textueller Einheiten zu manifestieren und also computerbasiert rezipierbar zu machen. Sie steht in Zusammenhang mit der Hypothese von der Externalisierung intertextueller (und – so ist hier zu ergänzen – intratextueller) Relationen mit Hilfe von Hypertexten (Sager 1997). Die Explikationshypothese ist eine abstrakte Hypothese, die konkretere Aussagekraft erst dadurch gewinnt, dass spezifiziert wird, was unter Kohäsion und Kohärenz genauer zu verstehen ist. Die Festlegung auf einen bestimmten Textbegriff mit seiner Definition des Kohärenzbegriffs führt zu einer charakteristischen Instanziierung der Explikationshypothese, derzufolge Links zur Explikation jenes Aspekts der Kohärenz textueller Einheiten herangezogen werden können, den dieser (hypertexttheoretisch entsprechend modifizierte) Kohärenzbegriff beschreibt. Dieser Zusammenhang wird ausführlich in Mehler 2005 erläutert.

Die Explikationshypothese schlägt eine Brücke zwischen Ausdrucks- und Inhaltsseite von Hyperlinks, und zwar mittels des Begriffs der Kohärenzrelation. Die Frage nach der Eigenart der hypertextuellen Kohärenz, die ausführlich in Storrer 2002 und 2003 erörtert wird, lässt diese Hypothese außer Acht. Jedoch kann je nach Ausgestaltung des Bedeutungsbegriffs unter anderem von struktural-semantischer, Referenz-semantischer, thematischer oder konzeptueller Ähnlichkeit der betroffenen Texte gesprochen werden. Betrachtet man dieser Vielfalt gegenüber einschlägige Verfahren zur Bemessung der inhaltlichen Ähnlichkeit von Texten, so fällt auf, dass sich diese im Wesentlichen an dem lexikalischen Material der Texte orientieren und höher geordnete Inhaltsbegriffe ebenso außer Acht lassen wie ihre Fundierung im Sinne des Kohärenzbegriffs. Inhaltliche Ähnlichkeit wird also weitgehend als sprachliche und insbesondere als lexikalische Ähnlichkeit von Texten beschrieben. Unabhängig von dieser Einschränkung ist aber festzuhalten, dass keines der Verfahren die Interpretationshoheit menschlicher Rezipienten in Frage stellt. Denn genau genommen unterbreiten diese Verfahren Informationsangebote, die erst kraft menschlicher Rezeptionsleistungen zum Wissensaufbau taugen. Trotz des nahe liegenden Hilfscharakters automatischer Verfahren zur Berechnung der inhaltlichen Ähnlichkeit von Texten soll dennoch möglichen Missverständnissen entgegengetreten werden:

- *Problematische Metaphern:* In einer namhaften Interpretation des Text Minings heißt es bei Hearst 1999, dass es eine Art des „Goldschürfens" darstelle, das „heretofore unknown", „never-before encountered information" aus Texten zu Tage fördere, und zwar über jene „realweltlichen" Zusammenhänge, welche die Texte thematisieren. Wiegand (2000:21) hält Metaphern dieser Art zu Recht entgegen, dass Texte keine Wissensträger sind, die Wissen über die „Wirklichkeit" übertragen; er negiert die Auffassung, wonach Wissen ein extrapersonaler „Kanalzustand" ist und hält dem entgegen, dass es „anhand von Inputdaten (besonders solcher sprachlicher Natur) kognitiv erarbeitet werden [muss], was nur möglich ist, wenn diese regelgerecht erzeugt wurden […]." Was also „heretofore unknown, never-before encountered information" ist, kann kein Textanalysesystem für sich entscheiden. Hierfür bedarf es stets der Rückbindung an menschliche Interpretationsprozesse. Medina-Mora et al. (1993:392) drücken dies wie folgt aus: „What is lost in the information perspective is the recognition that information in itself is uninteresting. Information is only useful because someone can do something with it, and we cannot define "do something" circularly as simply handling of more information. […] Business processes are implemented in information processes, just as information processes are implemented in material processes".
- *Die Offenheit der Informationssuche und die Korpusgebundenheit der Analyseergebnisse:* Ein weiteres Missverständnis betrifft die Fehleinschätzung der Dokumentsuche, welche durch die automatische Textanalyse und den Ähnlichkeitsvergleich von Dokumenten ermöglicht wird. So ist der Schluss naheliegenderweise falsch, dass aus dem Nicht-Nachweis eines Dokuments im Rahmen solcher Suchen seine Nicht-Existenz, zumindest aber seine Irrelevanz zu folgern ist, so als ob sich die Informationssuche in einer geschlossenen, vollständigen Welt bewegte. Die Ablehnung dieser Sichtweise steht in Zusammenhang mit der Einschätzung, dass automatisch gewonnene Analyseergebnisse in der Regel keinen Signifikanzanspruch in Bezug auf die jeweilige sprachliche Grundgesamtheit erheben (sollten). Was an automatisch explorierten Ähnlichkeitsurteilen auf andere Sprachgesamtheiten übertragbar ist, sind nicht die Urteile selbst, sondern das Verfahren zu ihrer Herleitung.
- *Die Unausweichlichkeit des Evaluationszirkels:* Ein weiteres Problem betrifft die Evaluation von Verfahren der automatischen Textanalyse, die bloß verfahrenstechnisch, nicht aber im Hinblick auf die mit ihr verbundene Interpretationsleistung automatisierbar ist. In unbeabsichtigter Distanz zum Begriff des *Text Minings* bringt dies der Begriff des anthropomorph überladen wirkenden Begriffs des *knowledge discovery in databases* (KDD; Fayyad et al. 1996) zum Ausdruck, indem er Interpretation und Evaluation der explorierten Daten an menschliche Leistungen bindet, die wiederum rekursiv auf jene Prozesse einwir-

ken, welche der Auswahl der zu analysierenden Daten zugrunde liegen (siehe hierzu Abbildung 1). Die Unabdingbarkeit einer solchen Rückbindung betrifft, so die hier vertretene Ansicht, jede Form der automatischen Textanalyse.

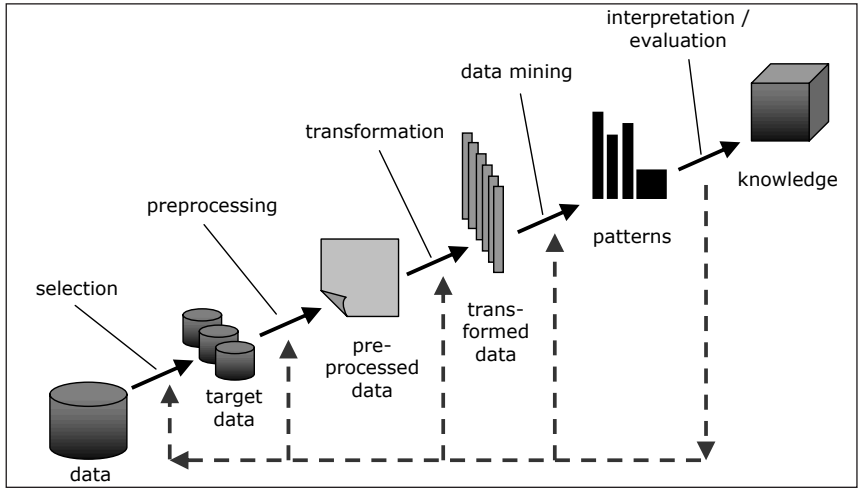

Abb. 1: Der KDD-Prozess nach Fayyad et al. (1996)

- *Teilautomatisierung:* Vollautomatische Textanalysen sind dieser Lesart zufolge unmöglich. Nichtsdestotrotz setzt Automatisierung Berechenbarkeit voraus und somit im vorliegenden Anwendungszusammenhang scheinbar also die algorithmische Durchdringung kognitiver Prozesse. Doch obgleich das Modelloriginal dieser Algorithmen kognitive Prozesse sein *können*, aber nicht müssen, gilt in Analogie zur Ablehnung anthropomorpher Deutungen der Daten, auf denen sie operieren, die Rückweisung der Sichtweise, wonach diese Algorithmen kognitive Prozesse realisierten. Als Prozessmodelle bilden sie vielmehr Simulationen, die ihren Modelloriginalen in der gleichen Weise zugeordnet sind wie jedes andere formale Modellkonstrukt. Hier besteht keine Konkurrenzsituation zwischen menschlicher Interpretation und maschineller Datenanalyse.
- *Der Zielkonflikt von Automatisierung und semantischer Fundierung:* Ein letzter kritischer Punkt betrifft die Tiefe der automatischen Textanalyse, inwieweit diese also nicht bloß textoberflächenstrukturelle, sondern zugleich semantische oder gar pragmatische Informationen auswertet. Darauf, dass es hier deutliche Grenzen gibt, die vielfach übersehen werden, weisen Cunningham et al. 1995 hin, und zwar unter Rekurs auf die syntaktische Analyse: „[...] by 1987 it was perhaps the largest DCG (Definite Clause Grammar) anywhere, designed to cover a linguistically well-motivated test set of sentences in English. Interpreted by a standard parser it was able to parse completely and uniquely virtually no sentence chosen randomly from a newspaper". Vor dem Hintergrund der mit solchen Versuchen verbundenen Erkenntnis rekurriert die Mehrzahl der anwendungsrelevanten Ansätze zur Textanalyse auf statistische, direkt anhand des beobachteten Textmaterials berechenbare Verfahren, unter (noch immer) weitgehender Ausklammerung linguistischen Wissens.

Diese kritischen Vorüberlegungen sollen nicht den falschen Schluss evozieren, automatische Textanalysen wären ein Widerspruch in sich selbst. Es geht vielmehr darum, den Schluss abzuwenden, dass vor dem Hintergrund ihrer offensichtlichen Erfolge der Eindruck entsteht, sie seien mehr als maschinelle Datenanalyse. Dieses Zurechtrücken bedeutet aber keine Abstufung der automatischen Textanalyse, sondern betont die Notwendigkeit ihrer zeichentheoretischen Einbettung, wie sie bestehende Verfahrensangebote vielfach vermissen lassen.

3. Automatische Textrelationierung

Der folgende Abschnitt erörtert Verfahren aus dem Bereich von Informatik und Computerlinguistik, die – ob nun in Kenntnis oder Unkenntnis der genannten Bedenken – auf die Aufgabe der automatischen Textanalyse zielen, wie sie als Voraussetzung für die automatische Verlinkung von Texten charakterisiert wurde. Trotz ihres formalen Charakters werden diese Verfahren nur informell skizziert. Ziel ist es, einen Überblick über den Verfahrensstand zu gewinnen, und zwar zu dem Zweck ihrer texttechnologischen Kritik.

3.1. Der methodische Rahmen

Vor dem Hintergrund der mit dem Internet gegebenen Distributionsmöglichkeiten ergibt sich das Dilemma, dass die Zahl online verfügbarer Dokumente auch im Wissenschaftsbereich rasant wächst (Lossau:2004), während der Zugriff auf diese Dokumente noch immer an die Möglichkeiten herkömmlicher Suchmaschinen gebunden ist (Göggler 2003). Dabei kommen primär Verfahren zum Einsatz, die sich an den Vorkommen von Such-Termen in Dokumenten orientieren und infolgedessen vergleichbare

Defizite erzielen, wie das klassische *Information Retrieval*. Ein Ausdruck dieses Defizits besteht darin, dass inhaltlich irrelevante Texte auch dann auf Suchanfragen hin nachgewiesen werden, wenn sie mindestens einen der Suchterme enthalten. Digitale Bibliotheken zielen demgegenüber auf einen informationellen Mehrwert, indem sie ihre Dokumentbestände *inhaltsorientiert* zugänglich machen (Kuhlen 1996; Fox/Sornil 1999; Lossau 2004). Ihr Erfolg ist eng an Fortschritte auf dem Gebiet der Erforschung von Verfahren zur automatischen Textanalyse gebunden. Das Spektrum dieser Verfahren reicht vom Parsing der Inhaltsstruktur von Texten (Hahn 1995) über die Zusammenfassung (Mani 2001) und Kategorisierung (Joachims 2002) ihrer Inhalte bis hin zum Nachspüren themengleicher Texte (Allan 2002a) und ihrer Verlinkung zu Hypertexten (Mehler 2004).

Die methodische Basis dieser Verfahren bildet ein *mathematisches Repräsentationsformat*, das wiederum als Voraussetzung für die geforderte *automatische* Verarbeitung gilt. Seine Auswahl ist durch den Zielkonflikt von automatischer Berechenbarkeit und adäquater Inhaltsrepräsentation gekennzeichnet, was sich am Beispiel des *Information Retrievals* demonstrieren lässt. Das hier vorwiegend eingesetzte *Vektorraummodell* (Salton 1989) repräsentiert nämlich Dokumente als Merkmalsvektoren, was zwar eine effiziente Verarbeitung der Dokumente garantiert, dafür aber die Repräsentation ihrer Inhalte an die Ausdruckskraft von Indextermen bindet. Was bedeutet das? Das bedeutet, dass Dokumente und Suchanfragen auf Vektoren abgebildet werden, deren Koordinaten den zuvor ausgewählten Indextermen entsprechen. Die Auswahl der Terme orientiert sich daran, wie gut sie Dokumente identifizieren und gleichzeitig diskriminieren, inwieweit sie also relevante und irrelevante Dokumente zu separieren erlauben, wenn diese nach dem Kriterium ermittelt werden, ob sie die Suchterme enthalten. Folgerichtig gelten Vertreter geschlossener Wortarten als schlechte Indexterme, da sie diese Eigenschaft nicht besitzen. Der Wert der Koordinate eines Dokumentvektors gibt somit an, inwieweit der durch die Koordinate repräsentierte Term diese Identifikations- und Diskriminanzfunktion für das Dokument besitzt. Dabei werden aus der Sicht eines Dokuments solche Terme hoch bewertet, die zur Konzentration in wenigen Dokumenten neigen, in denen sie jedoch häufig vorkommen. Was in diesem Sinne nicht indexiert wird, bleibt unauffindbar.

Das Vektorraummodell legt eine geometrische Interpretation des *Information Retrievals* nahe: Dokumente und Suchanfragen werden zunächst auf den Vektorraum projiziert. Durch den Vergleich ihrer Vektorrepräsentationen wird eine Rangfolge der Dokumente bezogen auf die Suchanfrage erstellt, welche die zunehmende Distanz von zugehörigem

Dokumentvektor und Vektor der Suchanfrage abbildet. Je größer dieser Abstand ist, desto geringer ist die Ähnlichkeit von Dokument und Anfrage und desto geringer ist die Wahrscheinlichkeit, dass das Dokument in das Suchergebnis aufgenommen wird. Da der Vektorvergleich auf dem Vergleich von Koordinatenwerten beruht, welche die Dokument-bezogene Relevanz von Indextermen wiedergeben, führt das Vektorraummodell letztlich einen gewichteten Vergleich des lexikalischen Materials von Dokumenten und Suchanfragen durch: Je höher die Zahl gemeinsamer Wörter und je größer ihre Indexierungsrelevanz, desto geringer ist der Abstand der zugehörigen Vektoren.

Im IR herrschen Interpretationen vor, wonach letztere Distanzwerte inhaltliche Ähnlichkeiten abbilden, was offenkundig falsch ist (Mehler 2004a). Tatsächlich findet das Vektorraummodell vor allem wegen seiner Einfachheit Anwendung. Die gegenwärtige Forschung zielt denn auch auf eine Ablösung dieses *bag-of-words*-Ansatzes, um seine mangelnde semantische Ausdruckskraft zu überwinden, freilich zu dem Preis einer höheren Berechnungskomplexität. In Analogie zur Unterscheidung von Textinhalt, -struktur und -funktion stehen hierfür mit der Inhalts-, Struktur- und Benutzermodellierung drei Bezugsgrößen bereit. Diese Analogie dient allein klassifikatorischen Zwecken, denn tatsächlich sind diese Ansätze linguistisch betrachtet vielfach unfundiert.

Inhaltsorientierte Retrievalmodelle

Mit dem latenten semantischen Indexieren (Dumais 1995) wurde ein algebraisches Retrievalmodell (Baeza-Yates/Ribeiro-Neto 1999) entwickelt, das sich an der Aufgabe orientiert, Suchanfragen und Dokumente auch dann als zusammengehörend zu identifizieren, wenn sie keine Terme gemeinsam haben, dafür aber ähnliche Inhalte behandeln. Anders als beim Vektorraummodell gelten dabei auch solche Dokumente als ähnlich, denen Indexterme mit ähnlichen Gebrauchsregularitäten gemeinsam sind. In einer Reihe von empirischen Untersuchungen konnte die prinzipielle Überlegenheit dieses Verfahrens gegenüber dem Vektorraummodell aufgezeigt werden (Dumais 1995), was die Frage nach seiner informationstheoretischen Basis aufwarf, für die mit der *Latent Semantic Analysis* (LSA) eine kognitionslinguistische Antwort geliefert wurde (Landauer/Dumais 1997). Auch wenn dieses Ergebnis in der Folgezeit anhand von Wortergänzungstests, Disambiguierungsaufgaben und Experimenten zu Prädikat-Argument-Strukturen (Schütze 1997; Kintsch 2001) bestätigt wurde, ist kritisch hervorzuheben, dass die LSA genauso wie das Vektorraummodell die *Struktur* von Dokumenten unberücksichtigt lässt.

Strukturierte Dokumente und linguistische Retrievalmodelle

Ausgehend von dem Konzept der *logischen Dokumentstruktur* (Power et al. 2003) identifizieren strukturorientierte Retrievalmodelle die (in der Regel mittels der *eXtensible Markup Language* ausgezeichneten) Komponenten von Dokumenten als zusätzliche Retrievaleinheiten (Lalmas/Ruthven 1998). Hierzu werden Dokumente als geordnete Hierarchien von Inhaltsobjekten beschrieben (Renear 1997), die unter anderem aus Sektionen, Paragraphen und Absätzen bestehen. Beim Retrieval strukturierter Dokumente geht es darum, inhaltlich relevante *Komponenten* eines Dokuments auch dann auffindbar zu machen, wenn das Dokument als Ganzes als irrelevant eingestuft wird. Ferner bietet die Dokumentgliederung einen zusätzlichen Bezugspunkt für die merkmalsorientierte Dokumentrepräsentation (Fuhr/Buckley 1991). In beiden Ansätzen ist die Identifikation der Dokumentstruktur an den Einsatz computerlinguistischer Verfahren gebunden (Mehler/Lobin 2004), welche die Dokumente aufgrund sprachlicher und inhaltlicher Kriterien segmentieren. Ein weiteres Anwendungsgebiet dieser Verfahren bildet die Vorverarbeitung von Dokumenten zur verbesserten Auswahl von Indextermen (Strzalkowski 1999). Dabei geht es über die klassische Grundformenreduktion hinaus um die Disambiguierung von Indextermen (Schütze 1997) wie auch um die Identifikation von Eigennamen und von Instanzen komplexer Ereignisschemata im Rahmen der Informationsextraktion (Lin 1998).

Im Kern rekurrieren diese Ansätze nach wie vor auf Merkmalsvektoren als grundlegendes Repräsentationsformat, wobei Dokumentstrukturen nur insoweit Berücksichtigung finden, als sie Retrievaleinheiten und deren Merkmale besser einzugrenzen erlauben. Die gleichzeitige Berechnung der Ähnlichkeit von Dokumenten aufgrund ihrer Struktur und ihres Inhalts wird ebenso vernachlässigt wie die dokumentübergreifende Exploration jenes Themennetzes, in das die Dokumente thematisch eingeordnet sind. Aus der Sicht digitaler Bibliotheken ist aber gerade der Rekurs auf die inhaltliche Vernetzung der Dokumente entscheidend, da sie mit der thematischen Vielfalt der Wissenschaftskommunikation und den Grundprinzipien ihrer Organisation in Zusammenhang steht. Folgerichtig erfordern digitale Bibliotheken ein Repräsentationsformat, das auf einem Modell dieses Themennetzes – sozusagen in Form eines *Semantic Web* (Fensel et al. 2003) der Wissenschaftskommunikation – aufbaut. Erst der Bezug auf ein solches Modell der Themenvernetzung erlaubt die Erbringung jenes informationellen Mehrwerts in Form eines inhaltsgeleiteten Retrievals, der für digitale Bibliotheken eingefordert wird.

Benutzermodellierung

Während inhalts- und strukturorientierte Modelle die Retrievalleistung durch die teils computerlinguistische Anreicherung von Dokumentmodellen zu verbessern suchen, ist es die Anpassung an die Informationsbedürfnisse und Nutzungsgewohnheiten der Informationssuchenden, auf die benutzerorientierte Retrievalansätze zielen. Das klassische Anwendungsgebiet dieser Ansätze bildet das Browsing (Kuhlen 1991) in Hypertexten. Hierzu werden *adaptive Hypertextsysteme* entwickelt, in denen Benutzermodelle dazu dienen, Rezipienten ausschließlich Links zu solchen Textmodulen verfügbar zu machen, die ihre Informationsbedürfnisse bedienen und ihr Vorwissen wie auch ihre Rezeptionsgewohnheiten berücksichtigen (Kuhlen 1994; Hammwöhner 1997).

Aus der Sicht digitaler Bibliotheken bilden die Struktur- und Inhaltsorientierung von Retrievalmodellen unerlässliche Kriterien zur Erzielung informationeller Mehrwerte gegenüber Suchmaschinen. Dabei ist die Auffassung leitend, dass die Modellierung thematischer Präferenzen von Informationssuchenden ein besseres Dokumentmodell voraussetzt, als es mit dem Vektorraummodell gegeben ist. Bevor also überhaupt die Benutzermodellierung als eine Aufgabe digitaler Bibliotheken aufgegriffen werden kann, ist zunächst das Problem einer verbesserten Dokumentmodellierung anzugehen, wobei Struktur- und Inhaltsmodellierung zu integrieren sind. Hiermit ist ein Ansatz gemeint, der über dokumentinterne Strukturen hinaus die thematische Vernetzung jener Inhalte berücksichtigt, welche die Dokumente behandeln. Das hierfür einschlägige Gebiet des *Topic Trackings* beruht unter anderem auf der Textkategorisierung, die nun vorbereitend erläutert wird.

3.2. Textkategorisierung

Als eine unmittelbare Möglichkeit zur inhaltsorientierten Strukturierung von Retrieval-Ergebnissen erweist sich die Textkategorisierung, mit deren Hilfe Texte auf vorgegebene Inhaltskategorien abgebildet werden. Ein WWW-basiertes Beispiel bildet die Suchmaschine *Vivísimo*, die Suchergebnisse in Form von Bäumen anordnet, deren Knoten Inhaltskategorien und deren Kanten Hyponymierelationen der Kategorien entsprechen. Die Textkategorisierung steht der explorativen Textklassifikation gegenüber, die auf dem Typ des *unüberwachten maschinellen Lernens* basiert und infolgedessen keine Vorgaben hinsichtlich Art und Zahl der zu explorierenden Klassen macht. Demgegenüber beruht die Textkategorisierung auf dem Typ des *überwachten* Lernens, und zwar unter Vorgabe der zu lernenden

Kategorien, deren Parameter in Trainingsphasen auf der Basis expertenseitig *vor*kategorisierter Textkorpora bestimmt und in Validierungsphasen verfeinert werden. Das zurzeit erfolgreichste Kategorisierungsverfahren bilden die *Support Vector Machines*, die von Joachims (siehe Joachims 2002) auf textuelle Einheiten übertragen wurden. Ihre Anwendung im Bereich Web-basierter Dokumente stößt auf das Polymorphieproblem (Mehler et al. 2004), das grob gesprochen besagt, dass Webpages Komponenten der Instanzen von Webgenres (Jakobs 2003) bilden und dabei vielfach mehrere Strukturtypen dieser Genres *gleichzeitig* manifestieren. Infolgedessen sind sie mit Mehrfachkategorisierungen verbunden, die dem Prinzip der im mathematischen Sinne *funktionalen* Kategorisierung entgegenstehen. Von dieser Problematik abgesehen besteht ein grundlegender Interpretationszirkel der Kategorisierung darin, dass sie wegen der Vorgabe der Kategorien und der Anpassung der Kategorisierungsparameter an diese Kategorien quasi zu einer Art Selbstbestätigung neigt, fernab von der Binnengliederung der Textinhalte und der Dynamik ihrer zeitlichen Entwicklung.

3.3. Thematische Ordnung von Dokumentmengen

Ansätze, die darauf zielen, Mengen von Dokumenten textoberflächenstruktur- oder inhaltsbasiert anzuordnen, lassen sich mit Hilfe folgender Kriterien klassifizieren:

1. Zum einen sind Ansätze danach zu unterscheiden, ob sie die Ähnlichkeit von Dokumenten aufgrund formaler oder inhaltlicher Merkmale bewerten. Erstere Gruppe von Ansätzen vergleicht die in den Dokumenten vorkommenden Indexterme und operiert daher *textgeleitet*.
2. Die zweite Gruppe von Ansätzen macht demgegenüber die Verknüpfung von Texten von den Inhalten abhängig, die sie behandeln, und arbeitet daher *inhaltsgeleitet*.

	TEXTGELEITET	INHALTSGELEITET
PAARWEISE TEXTVERKNÜPFUNG	textgeleitete Verknüpfung	inhaltsgeleitete Verknüpfung
STRUKTURBILDUNG OBERHALB PAARWEISE VERKNÜPFTER TEXTE	textgeleitete Strukturbildung	inhaltsgeleitete Strukturbildung

Tab. 1: Ansätze zur Relationierung von Dokumenten.

3. Zum anderen sind Ansätze danach zu unterscheiden, ob sie Dokumente ausschließlich paarweise vergleichen und gegebenenfalls verknüpfen oder in Abgrenzung hiervon auf die Bildung größerer Zusammenhangsstrukturen zielen, wie z. B. in Form chronologisch geordneter Dokumentketten.

Aus der Kombination dieser Kriterien resultieren – wie in Tabelle (1) dargestellt – vier Klassen von Ansätzen zur Strukturierung von Dokumentmengen, von denen text- bzw. inhaltsorientierte Ansätze zur paarweisen Dokumentverknüpfung in unmittelbarem Zusammenhang mit der Verwendung des Vektorraummodells zur automatischen Konversion von Texten in Hypertexte stehen (Agosti et al. 1996; Smeaton 1996; Allan 1997). Ihre Verfahrensweise ergibt sich aus dem zugrunde gelegten Retrievalmodell und dem hieraus ableitbaren Begriff der Dokumentähnlichkeit, was in Abschnitt 3.1 erläutert wurde. Es verbleiben somit zwei Gruppen von Ansätzen, die im Folgenden erörtert werden:

1. Bei der *textgeleiteten Strukturbildung* geht es darum, Dokumentmengen nach dem Kriterium gemeinsamer Indexterme zu ordnen. Während Green 1998 und Ferret 2002 zu diesem Zweck unter anderem auf das Vektorraummodell zurückgreifen, ist es das latente semantische Indexieren, das Chen/Czerwinski 1998 einsetzen. Allen diesen Ansätzen ist gemeinsam, dass sie die Verkettung von Dokumenten letztlich nicht an ihre Inhalte knüpfen, so dass wegen der Intransitivität der zugrunde gelegten Ähnlichkeitsrelation auch zusammenhangslose Texte derselben Dokumentkette oder -liste zugeordnet werden können.
2. Dieses Defizit wird bei der *inhaltsgeleiteten Strukturbildung* von Dokumentmengen im Prinzip aufgehoben, da hier alle Glieder einer Dokumentkette als Instanzen derselben übergeordneten Themenkategorie identifiziert werden. Anders ausgedrückt: Texte werden miteinander verknüpft, wenn sie Instanzen von Inhaltskategorien bilden, die ihrerseits zusammenhängen. Im einfachsten Fall besteht dieser Zusammenhang in der Identitätsrelation.

Auch diese Verfahrensgruppen sind mit Defiziten im Hinblick auf die inhaltsorientierte Ordnung von Dokumenten verbunden, was nun am Beispiel des *Document Routings* und des *Topic Trackings* erläutert wird.

3.3.1. Document Routing

Das *Document Routing* (DR) zielt darauf, eine Folge von Dokumenten auf eine vorgegebene Menge von Inhaltskategorien abzubilden und anschließend je Kategorie nach dem Kriterium abnehmender Relevanz zu ordnen.

Das DR, das Baeza-Yates/Ribeiro-Neto 1999 als eine Erweiterung des *Document Filterings* beschreiben, inkorporiert somit die Textkategorisierung als eine Teilaufgabe, weswegen die hierfür einschlägigen Verfahren zur Anwendung kommen (Sebastiani 2002). So stellen beispielsweise Guthrie et al. 1999 einen probabilistischen Routing-Ansatz basierend auf einem Bayeschen Klassifikator vor. Dieser Klassifikator bewertet unter anderem die Wahrscheinlichkeit, mit der aus der Beobachtung von Termen in einem Dokument auf seine Inhaltskategorie geschlossen werden kann. Diese Wahrscheinlichkeit wird durch Rekurs auf die Häufigkeit abgeschätzt, mit der diese Wörter in Texten bekannter Kategorie vorkommen.

Unabhängig von der Wahl des zugrunde liegenden Dokumentmodells – ob nun in Form des Vektorraummodells oder seiner Konkurrenten – besteht das Problem der Nutzbarmachung des Routings im Rahmen digitaler Bibliotheken darin, dass es die Kenntnis der verwendeten Inhaltskategorien voraussetzt. Dies steht im Gegensatz zur Dynamik der Wissenschaftskommunikation, in der sich immer neue Themenkategorien herausbilden, in Bezug aufeinander restrukturieren und auch wieder verschwinden, ohne auf das Konzept einer vorgegebenen endlichen Kategorienmenge reduzierbar zu sein. Das Routing erlaubt daher nicht die Bewältigung jener Dynamik, mit der digitale Bibliotheken konfrontiert sind.

3.3.2. Topic Tracking

Unter der Bezeichnung *Topic Detection and Tracking* (TDT) wurde eine Reihe von Ansätzen entwickelt, die auf die Segmentierung, Kategorisierung und chronologische Ordnung von Dokumentströmen in Bezug auf die in ihnen thematisierten Ereignisse zielen. Gegenüber dem Routing rekurriert das Tracking auf einen ereignisorientierten Themenbegriff, wofür das Nachspüren von Pressetexten zu derselben Ereignisfolge als Paradebeispiel gilt. Dabei werden fünf Teilaufgaben unterschieden (Allan 2002a): Auf die *Story Segmentation* anhand von Segmentmarkern folgt die *First Story Detection*, das heißt die Erkennung neuer Ereignisse mit Hilfe von Verfahren der automatischen Klassifikation (Bock 1994), ferner die *Cluster Detection* (das heißt das ereignisorientierte Clustern der zuvor erkannten Segmente) sowie das *Topic Tracking* im engeren Sinne und schließlich die Verkettung der geclusterten Segmente im Rahmen der *Story Link Detection*. Methodisch betrachtet hebt sich das TDT vom Document Routing dadurch ab, dass es Verfahren der automatischen Klassifikation inkorporiert und also keine Vorgabe von Inhaltskategorien voraussetzt. In der Literatur finden sich hierzu Vektorraummodell-basierte Ansätze (Carthy/Smeaton 2000; Allan et al. 2002; Ji/Zha 2003) ebenso, wie probabilistische Ansätze, die das Tracking als einen Markov-Prozess beschreiben (Blei/Moreno 2001; Yamron et al. 2002), wobei die chronologische Ordnung einen weiteren Bezugspunkt für das Tracking bietet (Dalamagas/Dunlop 1997).

Das TDT hebt sich dadurch vom Information Retrieval ab, dass es auf die inhaltliche Ordnung von Dokumentmengen zielt. Es identifiziert immer dann ein neues Thema, wenn das Inputdokument den bereits detektierten *topics* nicht trennscharf zuzuordnen ist. Dabei bleibt allerdings ungeklärt, in welchem Verhältnis die neue Themeneinheit zu den bereits verfolgten *topics* steht, ob nun in der Funktion einer thematischen Fortsetzung, eines Abschlusses oder eines Teilthemas. Das Tracking reduziert sich somit auf die Verkettung von Einheiten der Dokumentebene, *ohne* die Vernetzung der darüber liegenden Themenebene abzubilden.

In eine andere Richtung der themengeleiteten Strukturierung weist der konnektionistische Ansatz der *Hierarchical Self-Organizing feature Maps* (HSOM). Er basiert auf einer geometrischen Interpretation von Ähnlichkeitsrelationen, derzufolge ähnliche Dokumente auf benachbarte, unähnliche hingegen auf distante Kartenregionen abgebildet werden. Im Falle von *Feature Maps* soll auf diese Weise eine zweidimensionale topologische Ordnung der Dokumente gewonnen werden. HSOMs gehen demgegenüber von einer hierarchischen Organisation der Themenkategorien aus. Hierzu erweitern sie das Konzept der *Self-Organizing Feature Map* (Kohonen 2001), indem sie die Merkmalskarten verschachteln, so dass bis zu einer vorgegebenen Schachtelungstiefe jedem Knoten der betrachteten Ebene genau eine Karte der darunter liegenden Ebene zugeordnet ist. Die thematische Klassifikation erfolgt in diesem Ansatz *top down*, wobei je Ebene der Themenknoten mit dem höchsten Aktivierungsniveau festlegt, durch welche Teilkarte der dominierten Ebene das Inputdokument weiterverarbeitet wird. Hierdurch werden Eingabemuster auf Folgen zunehmend feiner auflösender Inhaltskategorien abgebildet (Merkl 2000). Freilich wird dabei die Zahl der Ebenen ebenso vorgegeben wie die Zahl der Themenknoten je Teilkarte. Somit ist der HSOM-Ansatz mit vergleichbaren Defiziten verknüpft, wie das Document Routing, wenn es darum geht, die Dynamik der Wissenschaftskommunikation und also das Aufkommen, Restrukturieren und Verschwinden von Themenkategorien abzubilden.

Der Merkmalskartenansatz ist ebenso wie das TDT aus der Sicht der Anforderung nach themengeleiteter Strukturierung von Dokumentmengen verbesserungswürdig. Denn während der Kartenansatz die Strukturierung thematischer Einheiten auf das Konzept der Themenhierarchie

unter Vorgabe der Themenebenen reduziert, beschränkt das TDT die Strukturbildung auf das Konzept der listenförmigen Verkettung. Demgegenüber wird ein Ansatz benötigt, der das Konzept der hierarchischen Anordnung mit dem explorativen Charakter des TDT verbindet.

3.3.3. Kohäsionsbäume

In diese Richtung zielt ein Ansatz, der auf der Hypothese beruht, dass Texte und ihre Komponenten ebenso wie Wörter eine strukturelle Bedeutung besitzen, auf deren Grundlage ihre *semantische Ähnlichkeit* beurteilt werden kann. Um den für diese Aufgabenstellung charakteristischen Zielkonflikt von automatischer und zugleich inhaltsbasierter Textanalyse zu lösen, rekurriert er auf einen Algorithmus, der die inhaltliche Ähnlichkeit von Texten berechnet, ohne – wie das *Latente Semantische Indexieren* (LSI; Dumais 1995) – auf faktorenanalytische Konzepte zurückzugreifen (Mehler 2001). Dieser Ansatz erzielt gegenüber dem LSI eine größere computerlinguistische Transparenz, indem die Teilschritte des Konstruktionsalgorithmus zur Repräsentation von Textinhalten als Modelle syntagmatischer und paradigmatischer Prozesse identifizierbar sind, jedoch nicht nur auf Wortebene (Rieger 1989), sondern zugleich auf Textebene. Der Konstruktionsalgorithmus orientiert sich an dem Kriterium, die semantische Ähnlichkeit von Texten auch dann vorherzusagen, wenn sie nur wenige oder im Extremfall keine oberflächenstrukturellen Gemeinsamkeiten aufweisen, dafür aber ähnliche Inhalte behandeln. Folglich steht dieses Verfahren in der Tradition inhaltsbasierter Retrievalmodelle. Es geht jedoch insofern über konkurrierende Ansätze hinaus, als es die inhaltsbasierte Analyse von Texten mit ihrer Verkettung zu kohäsiven Textfolgen verbindet. Den Ausgangspunkt hierfür bildet eine Kritik der Orientierung an *Ergebnislisten* als grundlegende Form für die Strukturierung von Textmengen.

Document Routing und Topic Tracking zielen auf die listenförmige Anordnung themengleicher Dokumente. Inhaltsbasierte Relationen lassen jedoch eine Strukturierung in Form von thematischen Progressionen und Verzweigungen erkennen, welche Listen nicht erfassen. Listenförmige Anordnungen haben zur Folge, dass inhaltsverschiedene Dokumente aufgrund ihrer Ähnlichkeit zur Suchanfrage benachbarte Rangplätze innerhalb der Ergebnisliste einnehmen und infolgedessen konsekutiv rezipiert werden können. Listenbasierte Verfahren lassen daher jede Übersicht über die Binnengliederung der von den Dokumenten behandelten Themen außer Acht. Im Gegensatz hierzu zielt der vorliegende Ansatz auf die thematische Gliederung von Texten entlang der von ihnen

behandelten Themen, und zwar auf der Basis linguistisch fundierter und automatisch berechenbarer *Kohäsionsbäume* (Mehler 2002). Sie dienen dem Ziel, dieselbe Dokumentmenge aus variabler thematischer Perspektive zu traversieren, wobei ihre Wurzel jenen Dokumenten entspricht, welche das Informationsbedürfnis des jeweiligen Informationssuchenden eingrenzen, während ihre Pfade die dokumentbasierte Entfaltung jener Themen zum Ausdruck bringen, welche im zugrunde liegenden Textkorpus behandelt werden. Die Berechnung von Kohäsionsbäumen basiert auf der folgenden Kriterienliste:

- *Kettenbildung:* Thematisch homogene Dokumente bilden die Knoten desselben Pfads, so dass ausgehend vom Wurzelknoten jeder Pfad die Entfaltung eines oder mehrerer, jedoch zusammengehöriger Themen abbildet.
- *Themenwechsel:* Mit wachsender Länge eines Pfads (gemessen an der Zahl seiner Knoten) steigt die Wahrscheinlichkeit, dass nachfolgende Knoten einen latenten Themenwechsel realisieren.
- *Verzweigung:* Thematisch mehrdeutige Dokumente bilden die Verzweigungspunkte von Pfaden, die je einen der thematischen Aspekte des mehrdeutigen Dokuments entfalten.
- *Interaktivität:* Informationssuchende können zur Laufzeit die Perspektive wechseln, unter der die Dokumentkollektion hierarchisch geordnet wird, und so die Erzeugung eines neuen Kohäsionsbaums mit neuer thematischer Wurzel veranlassen.

Kohäsionsbäume werden Informationssuchenden in Form von Hypertexten präsentiert. Empirische Experimente zeigen, dass dabei Verknüpfungsfehler entstehen, die im Rahmen letzterer Kriterienliste nicht auszuräumen sind (Mehler 2002a). Ein wesentlicher Grund hierfür besteht darin, dass nach wie vor die ähnlichkeitsbasierte Assoziation als grundlegendes Organisationsprinzip fungiert, auch wenn sie durch Bewertung der Beziehungen indirekt verknüpfter Texte korrigiert wird. Nichtsdestotrotz verzichten Kohäsionsbäume auf eine Fundierung der Textvernetzung auf der Basis des zugrunde liegenden Themennetzes, was als Hauptfehlerquelle zu identifizieren ist. Was also benötigt wird, ist eine Verbindung des Prinzips eines *Semantic Webs*, dessen Themenkategorien jedoch nicht vorgegeben, sondern automatisch exploriert werden, mit dem Prinzip der automatischen strukturorientierten Textvernetzung, wie sie Kohäsionsbäume realisieren, und zwar nicht nur auf der Ebene der externen, sondern auch der internen Textstrukturierung. In letzterem Falle ist es insbesondere die logische Dokumentgliederung in Sätze, Absätze,

Sektionen etc., welche als zusätzliches Strukturmerkmal für die Textvernetzung bereitsteht. Inwieweit dabei Register und Genre, mithin also Diskursthemen- und -handlungstypen zur automatischen Textvernetzung herangezogen werden können, verdeutlicht Mehler 2002b und 2005.

4. Ausblick

Die vorangehenden Abschnitte haben Verfahren der Textverlinkung thematisiert, wie sie im Bereich digitaler Bibliotheken Verwendung finden (könnten). Über diese Verfahren hinaus stellt sich die grundlegende Frage, welche Richtung die Entwicklung digitaler Bibliotheken einschlagen könnte, um dem Anspruch nach informationellen Mehrwerten (Kuhlen 1994) gegenüber bestehenden Suchmaschinen-Technologien einzulösen. Hierzu sollen abschließend zwei Entwicklungslinien skizziert werden:

- *Kapselung der terminologischen Variabilität der Wissenschaftskommunikation:* Dieselben Themen werden vielfach unter Rückgriff auf verschiedene Terminologien behandelt. Diese Variabilität ist unter anderem eine Funktion der Kovariation wissenschaftlicher Themen und Terminologien und also *zeitbedingt*. Infolgedessen ist die diachrone, aber auch synchrone Variabilität von Termen zu berücksichtigen, deren Rolle als Indexterme über Repräsentationen jener Themen zu ermitteln ist, zu deren veränderlichen Manifestation sie beitragen. In diesem Sinne sollten digitale Bibliotheken inhaltlich relevante Dokumente auch dann auffindbar machen, wenn sie in terminologischen Varianten verfasst sind. Hierzu sind Kontextvariablen in die Dokumentmodellierung zu integrieren, welche das veränderliche Verhältnis von Thema und Terminologie reflektieren.
- *Kapselung der medialen Variabilität:* Dokumente sollten auch dann über dieselbe skalierbare Schnittstelle recherchierbar bleiben, wenn sie von verschiedenen Bibliotheken unter Verwendung verschiedener Informationssysteme in verschiedenen Formaten verwaltet werden (siehe auch Lossau 2004).

Ein Bibliotheksverband, der eine digitale Bibliothek basierend auf diesen Prinzipien realisiert, würde es seinen Benutzern erlauben, dieselben Dokumente in stets erneuerbare, evolvierende Themenzusammenhänge zu stellen und so nach den neuesten, stets aktualisierten Suchkriterien auffindbar zu halten. Eine thematisch spezialisierte Bibliothek als Mitglied dieses Verbands wäre in die Lage zu versetzen, auch solche Angebote recherchierbar zu machen, die gar nicht zu ihrem Bestand zählen. Sie erhielte

gewissermaßen eine an ihre thematische Spezialisierung angepasste Suchmaschine zur inhaltsbasierten Recherche desjenigen Dokumentangebots, das sie idealerweise anböte, aufgrund vielfacher Restriktionen jedoch nicht anzubieten in der Lage ist. Die Texttechnologie, der sich der vorliegende Beitrag zuspricht, ist eine junge wissenschaftliche Disziplin, welche einen Beitrag zur Abmilderung dieser Restriktionen leisten könnte.

Literatur

Agosti, M., F. Crestani und M. Melucci (1996): Design and implementation of a tool for the automatic construction of hypertexts for information retrieval. Information Processing & Management 33(4), S. 459-476

Allan, J. (1997): Building hypertext using information retrieval. Information Processing & Management 33(2), S. 145-159

Allan, J. (2002a): Introduction to topic detection and tracking. In: Allan (2002b), S. 1-16

Allan, J. (Ed.) (2002b): Topic Detection and Tracking. Event-based Information Organization. Boston: Kluwer

Allan, J., V. Lavrenko und R. Swan (2002): Explorations within topic tracking and detection. In: Allan (2002b), S. 197-224

Baeza-Yates, R. und B. Ribeiro-Neto (Hg.) (1999): Modern Information Retrieval. Reading: Addison-Wesley

Blei, D. M. und P. J. Moreno (2001, September): Topic segmentation with an aspect hidden markov model. In: Proc. of the 24th Annual Intern. SIGIR Conf. on Research and Development in IR, New York: ACM Press

Bock, H. H. (1994): Classification and clustering: Problems for the future. In: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand und B. Burtschy (Hg.), New Approaches in Classification and Data Analysis. Berlin u.a.: Springer, S. 3-24

Brinker, K. (1992): Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. Berlin: Erich Schmidt

Carthy, J. und A. F. Smeaton (2000): The design of a topic tracking system. In: Proc. of the 22nd Annual Colloq. on IR Research, Cambridge: The IR Specialist Group of the British Computer Society

Chen, C. und M. Czerwinski (1998): From latent semantics to spatial hypertext: An integrated approach. In: Proc. of 9th Conf. on Hypertext and Hypermedia. New York: ACM Press, S. 77-86

Cunningham, H./ Gaizauskas, R. G./ Wilks, Y. (1995): A General Architecture for Text Engineering (GATE) – a new approach to Language Engineering. Technical Report CS-95-21: Department of Computer Science, University of Sheffield

Dalamagas, T. und M. D. Dunlop (1997): Automatic construction of news hypertext. In: Proc. of the HIM '97. Konstanz: Universitätsverlag, S. 265-278

Dumais, S. T. (1995): Latent semantic indexing (LSI): Trec-3 report. In: Overview of the 3rd Text Retrieval Conference (TREC-3). Gaithersburg: NIST, S. 219-230

El-Beltagy, S. R., W. Hall, D. D. Roure und L. Carr (2001): Linking in context. In: Proc. of the 12th Conference on Hypertext and Hypermedia. New York: ACM Press, S. 151-160

Fayyad, U./ Piatetsky-Shapiro, G./ Smyth, P. (1996): The KDD Process for Extracting Useful Knowledge From Volumes of Data. In: Communications of the ACM 39(11), S. 27-34

Fensel, D./ Hendler, J./ Lieberman, H. and W. Wahlster (2003): Spinning the Semantic Web. Bringing the World Wide Web to Its Full Potential. Cambridge: MIT Press

Ferret, O. (2002): Using collocations for topic segmentation and link detection. In: Proc. of the 19th Int. Conf. on Computational Linguistics, San Francisco: Morgan Kaufmann, S. 260-266

Fox, E. A. und O. Sornil (1999): Digital libraries. In: R. Baeza-Yates und B. Ribeiro-Neto (Hg.), Modern Information Retrieval. New York: Addison-Wesley, S. 415-432

Fuhr, N. und C. Buckley (1991): A probabilistic learning approach for document indexing. ACM Transactions on Information Systems 9(3), S. 223-248

Göggler, M. (2003): Suchmaschinen im Internet. Berlin u.a.: Springer

Green, S. J. (1998): Automated link generation: Can we do better than term repetition? Computer Networks and ISDN Systems 30(1-7), 75-84

Guthrie, L., J. Guthrie und J. Leistensnider (1999): Document classification and routing. In: T. Strzalkowski, Natural Language Information Retrieval. Dordrecht: Kluwer, S. 289-310

Hahn, U. (1995): Distributed text structure parsing - computing thematic progressions in expository texts. In: G. Rickheit und C. Habel (Hg.), Focus and Coherence in Discourse Processing. Berlin u.a.: de Gruyter, S. 214-250

Hammwöhner, R. (1997): Offene Hypertextsysteme: das Konstanzer Hypertextsystem (KHS) im wissenschaftlichen und technischen Kontext. Konstanz: Universitätsverlag

Hearst, M. A. (1999): Untangling Text Data Mining. In: Proc. of the 37th Annual Meeting of the Association for Computational Linguistics

Jakobs, E.-M. (2003): Hypertextsorten. In: Zeitschrift für germanistische Linguistik 31(2), S. 232-252

Ji, X. und H. Zha (2003): Extracting shared topics of multiple documents. In: Proc. 7th Pacific-Asia Conf. on Adv. in Knowledge Discovery and Data Mining. Berlin u.a.: Springer, S. 100-110

Joachims, T. (2002): Learning to classify text using support vector machines. Boston: Kluwer.

Kintsch, W. (2001): Predication. Cognitive Science 25, S. 173-202

Kobayashi, M./ Takeda, K. (2000): Information Retrieval on the Web. In: ACM Computing Surveys 32(2), S. 144-173

Kohonen, T. (2001): Self-Organizing Maps. Berlin u.a.: Springer

Kuhlen, R. (1991): Hypertext: ein nichtlineares Medium zwischen Buch und Wissensbank. Berlin u.a.: Springer

Kuhlen, R. (1994): Annäherung an Informationsutopien über offene Hypertextsysteme. In: R. Wille und M. Zickwolff (Hg.), Begriffliche Wissensverarbeitung: Grundlagen und Aufgaben. Mannheim: BI, S. 191-224

Kuhlen, R. (1996): Zur Virtualisierung von Bibliotheken und Büchern. In: D. Matejovski und F. Kittler (Hg.), Literatur im Informationszeitalter. Frankfurt: Campus, S. 112-142

Lalmas, M. und I. Ruthven (1998): Representing and retrieving structured documents using the dempster-shafer theory of evidence. Journal of Documentation 54(5), S. 529-565

Landauer, T. K. und S. T. Dumais (1997): A solution to plato's problem. Psychological Review 104(2), S. 211-240

Lawrence, S./ Giles, C. L./ Bollacker, K. (1999): Digital Libraries and Autonomous Citation Indexing. In: IEEE Computer 32(6), S. 67-71

Lin, D. (1998): Using collocation statistics in information extraction. In: Proceedings of the Seventh Message Understanding Conference 1998, MUC-7

Lobin, H. und Lemnitzer, L. (Hg.): Texttechnologie. Perspektiven und Anwendungen. Tübingen: Stauffenburg

Lossau, N. (2004): Search engine technology and digital libraries: Libraries need to discover the academic internet. D-Lib Magazine 10(6)

Mani, I. (2001): Automatic Summarization. Amsterdam: John Benjamins

Medina-Mora, R./ Winograd, T./Flores, R and F. Flores (1993): The Action Workflow Approach to Workflow Management Technology. In: The Information Society 9(4), S. 391-404

Mehler, A. (2001): Textbedeutung. Zur prozeduralen Analyse und Repräsentation struktureller Ähnlichkeiten von Texten. Frankfurt a. M.: Peter Lang

Mehler, A. (2002a): Hierarchical Orderings of Textual Units. In: Proc. of the 19th Int. Conf. on Computational Linguistics, San Francisco: Morgan Kaufmann, S. 646-652

Mehler, A. (2002b): Components of a Model of Context-Sensitive Hypertexts. In: Journal of Universal Computer Science (J.UCS) 8(10), S. 924-943

Mehler, A. (2003): Ein Kompositionalitätsprinzip für numerische Textsemantiken. In: LDV Forum 18(1-2), S. 321-337

Mehler, A. (2004a): Textmodellierung: Mehrstufige Modellierung generischer Bausteine der Textähnlichkeitsmessung. In: Mehler/Lobin (2004b), S. 101-120

Mehler, A. (2005): Ansätze zur textlinguistischen Fundierung der teilautomatischen Generierung von Hypertexten. Erscheint in: Sprache und Datenverarbeitung

Mehler, A./ M. Dehmer/ R. Gleim (2004): Towards Logical Hypertext Structure - A Graph-Theoretic Perspective. In: Proc. of the 4th International Workshop on Innovative Internet Computing Systems. Berlin u.a.: Springer

Mehler, Al./ Lobin, H. (2004a): Aspekte der linguistischen Informationsmodellierung. In: Mehler/Lobin (2004b), S. 1-21

Mehler, A./ Lobin, H. (Hg.) (2004b): Automatische Textanalyse: Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte. Wiesbaden: Verlag für Sozialwissenschaften.

Merkl, D. (2000): Text data mining. In: R. Dale, H. Moisl und H. Somers (Hg.), Handbook of Natural Language Processing. New York: Dekker, S. 889-903

Power, R., D. Scott und N. Bouayad-Agha (2003): Document structure. Computational Linguistics 29(2), S. 211-260

Rieger, B. (1989): Unscharfe Semantik: die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten. Frankfurt a. M.: Peter Lang

Renear, A. (1997): Out of praxis: Three (meta)theories of textuality. In: K. Sutherland (Hg.), Electronic Text. Investigations in Method and Theory. Oxford: Clarendon Press, S. 107-126

Sager, S. F. (1997): Intertextualität und die Interaktivität von Hypertexten. In: J. Klein und U. Fix (Hg.), Textbeziehungen: linguistische und literaturwissenschaftliche Beiträge zur Intertextualität. Tübingen: Stauffenburg, S. 109-123

Salton, G. (1989): Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading: Addison Wesley

Schütze, H. (1997): Ambiguity Resolution in Language Learning: Computational and Cognitive Models. Stanford: CSLI Publications

Sebastiani, F. (2002): Machine learning in automated text categorization. ACM Computing Surveys 34(1), S. 1-47

Smeaton, A. F. (1996): Building hypertext under the influence of topology metrics. In: Proc. of the Int. Workshop on Hypermedia Design. Berlin u.a.: Springer, S. 105-106

Storrer, A. (2002): Coherence in text and hypertext. Document Design 3(2), S. 156-168

Storrer, A. (2003): Text und Hypertext. In: H. Lobin und L. Lemnitzer (Hg.), Texttechnologie. Perspektiven und Anwendungen. Tübingen: Stauffenburg

Strzalkowski, T. (Ed.) (1999): Natural Language Information Retrieval. Dordrecht: Kluwer

Wiegand, H. E. (2000): Wissen, Wissensrepräsentation und Printwörterbüchern. In: Proc. of the 9th EURALEX International Congress, Stuttgart: IMS, S. 15-38

Yamron, J. P., L. Gillick, P. van Mulbregt und S. Knecht (2002): Statistical models of topical content. In: Allan (2002b), S. 115-134