Automatisches Klassifizieren

Verfahren zur Erschliessung elektronischer Dokumente

Master's Thesis
Zusatzstudiengang Bibliotheks- und Informationswissenschaft
Fakultät für Informations- und Kommunikationswissenschaften
Fachhochschule Köln

vorgelegt von:

Dr. Otto Oberhauser
Türkenstrasse 12, A 1090 Wien, Österreich
Matr. Nr.: 11035223

am 27. Juli 2004

Betreuer: Prof. Dipl.-Math. Winfried Gödert
Prof. Dr. Achim Oßwald

Abstract

Automatisches Klassifizieren von Textdokumenten bedeutet die maschinelle Zuordnung jeweils einer oder mehrerer Notationen eines vorgegebenen Klassifikationssystems zu natürlichsprachlichen Texten mithilfe eines geeigneten Algorithmus. In der vorliegenden Arbeit wird in Form einer umfassenden Literaturstudie ein aktueller Kenntnisstand zu den Einsatzmöglichkeiten des automatischen Klassifizierens für die sachliche Erschliessung von elektronischen Dokumenten, insbesondere von Web-Ressourcen, erarbeitet. Dies betrifft zum einen den methodischen Aspekt und zum anderen die in relevanten Projekten und Anwendungen gewonnenen Erfahrungen. In methodischer Hinsicht gelten heute statistische Verfahren, die auf dem maschinellen Lernen basieren und auf der Grundlage bereits klassifizierter Beispieldokumente ein Modell – einen "Klassifikator" – erstellen, das zur Klassifizierung neuer Dokumente verwendet werden kann, als "state-of-the-art". Die vier in den 1990er Jahren an den Universitäten Lund, Wolverhampton und Oldenburg sowie bei OCLC (Dublin, OH) durchgeführten "grossen" Projekte zum automatischen Klassifizieren von Web-Ressourcen, die in dieser Arbeit ausführlich analysiert werden, arbeiteten allerdings noch mit einfacheren bzw. älteren methodischen Ansätzen. Diese Projekte bedeuten insbesondere aufgrund ihrer Verwendung etablierter bibliothekarischer Klassifikationssysteme einen wichtigen Erfahrungsgewinn, selbst wenn sie bisher nicht zu permanenten und qualitativ zufriedenstellenden Diensten für die Erschliessung elektronischer Ressourcen geführt haben. Die Analyse der weiteren einschlägigen Anwendungen und Projekte lässt erkennen, dass derzeit in den Bereichen Patent- und Mediendokumentation die aktivsten Bestrebungen bestehen, Systeme für die automatische klassifikatorische Erschliessung elektronischer Dokumente im laufenden operativen Betrieb einzusetzen. Dabei dominieren jedoch halbautomatische Systeme, die menschliche Bearbeiter durch Klassifizierungsvorschläge unterstützen, da die gegenwärtig erreichbare Klassifizierungsgüte für eine Vollautomatisierung meist noch nicht ausreicht. Weitere interessante Anwendungen und Projekte finden sich im Bereich von Web-Portalen, Suchmaschinen und (kommerziellen) Informationsdiensten, während sich etwa im Bibliothekswesen kaum nennenswertes Interesse an einer automatischen Klassifizierung von Büchern bzw. bibliographischen Datensätzen registrieren lässt. Die Studie schliesst mit einer Diskussion der wichtigsten Projekte und Anwendungen sowie einiger im Zusammenhang mit dem automatischen Klassifizieren relevanter Fragestellungen und Themen.

Inhaltsverzeichnis

Danksagung

Für die Vergabe des Themas und die Betreuung der vorliegenden Masterarbeit bin ich Herrn Prof. Dipl.-Math. Winfried Gödert, FH Köln, zu Dank verpflichtet. Herrn Prof. Dr. Achim Oßwald danke ich für die Übernahme der Zweitbetreuung sowie für Rat und Hilfe während meines Studiums an der FH Köln.

Ausserdem danke ich folgenden Personen und Institutionen für ihre Unterstützung bei der Suche und Beschaffung von Informationen und Literatur: Gerry McKiernan (Ames, IA); Gerhard Möller-Schwing (Bremen); Prue Deacon (Canberra); Diane Vizine-Goetz (Dublin, OH); Eva Bertha, Elisabeth Böllmann (Graz); Hans-Joachim Bentz (Hildesheim); Heinz Hauffe, Georg Stern-Erlebach, Martin Wieser, Eveline Pipp (Innsbruck); Sr. Angelika und Sr. Helene (Kloster Kirchberg am Wechsel); Winfried Gödert, Klaus Lepsky, Achim Oßwald (Köln); Adalbert Kirchgäßner (Konstanz); Heinz-Dieter Knöll (Lüneburg); Traugott Koch (Lund); Institut für Rundfunktechnik GmbH (München); Bernd Diekmann (Oldenburg); Harald H. Zimmermann (Saarbrücken); Peter Pilsl (Salzburg); Christian Authried, Johann Brandauer, Martin Hekele, Helmuth Höbarth; Hans Hrusa, Günter Kindl, Peter Klien, Silvia Köpf, Peter Kubalek, Bernhard Kurz, Josef Labner, Angelika Laburda, Günther Müller, Günter Olensky, Wolfram Seidler, Karl Stebegg, Robert Würzl, Harald Mittermann (Wien).

Wien, im Juli 2004.

Abbildungs- und Tabellenverzeichnis

Abbildungen:

Tabellen:

Verwendete Abkürzungen und Akronyme

Abb.	Abbildung(en)
ACE	Automatic Classification Engine
ACM	Association for Computing Machinery
ACN	Automatic Classification Numbering [System]
ACS	Automatic Classification System
AE	"All" Engineering [Datenbank]
AUTCS	[UDC Number] Automatic Combination System
BIS	Bibliotheks- und Informationssystem (Univ. Oldenburg)
BNB	British National Bibliography
bspw.	beispielsweise
BUBL	Bulletin Board for Libraries
BVB	Bibliotheksverbund Bayern
bzw.	beziehungsweise
CC	Colon Classification
CSS	Cascading Style Sheet
DDC	Dewey Decimal Classification
DESIRE	Development of a European Service for Information on Research and Education
DFG	Deutsche Forschungsgemeinschaft
d.h.	das heisst
Dipl.-Arb.	Diplomarbeit
DocCat	Document Categorizer
DutchESS	Dutch Electronic Subject Service
ECLA	European Classification
EELS	Engineering Electronic Library, Sweden
Ei	Engineering Information, Inc.
EPA	Europäisches Patentamt (Rijswijk, Niederlande)
ESS	Editorial Support System [OCLC]
et al.	und andere
ETH	Eidgenössische Technische Hochschule
EU	Europäische Union
FH	Fachhochschule
G+J	Gruner + Jahr (Hamburg)
GBV	Gemeinsamer Bibliotheksverbund (Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Sachsen-Anhalt, Schleswig-Holstein, Thüringen)
GERHARD	German Harvest Automated Retrieval and Directory
Hrsg.	Herausgeber

HS	Hochschule
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
ibid.	ibidem, ebenda
ILRT	Institute for Learning and Research Technology (Univ. Bristol)
INPI	Institut National de la Propriété Industrielle (Paris)
IPC	International Patent Classification
IPCC	Industrial Property Cooperation Center (Japan)
IR	Information Retrieval
ISIV	Institut für Semantische Informationsverarbeitung (Univ. Osnabrück)
k-NN	*k* Nearest Neighbors-Verfahren
Kpt.	Kapitel
LCC	Library of Congress Classification
LCSA	Library of Congress Subject Authority [file]
LCSH	Library of Congress Subject Headings
LIS	Library and Information Science
Liz.-Arb.	Lizenziatsarbeit
LoC	Library of Congress (Washington, DC)
L/U	Last Update, last modified, letzte Aktualisierung
M.-Arb.	Masterarbeit
m.E.	meines Erachtens, [nach] meinem Erachten
m.W.	meines Wissens, , [nach] meinem Wissen
MAB	Maschinelles Austauschformat für Bibliotheken
MARC	Machine-Readable Cataloging [bibliographisches Austauschformat]
MeSH	Medical Subject Headings
NDC	Nippon Decimal Classification
NORDINFO	Nordic Council for Scientific Information and Research Libraries
o.a.	oben angeführt
o.J., o.O.	ohne Jahresangabe, ohne Ortsangabe
OCLC	Online Computer Library Center, Inc. (Dublin, OH)
OFFIS	Oldenburger Forschungs- und Entwicklungsinstitut für Informatikwerkzeuge und -systeme
OMNI	Organising Medical Networked Information
OPAC	Online Public Access Catalogue
p.	pagina(e), page(s)
PDF	Portable Document Format
PLSA	Probabilistic Latent Semantic Analysis
Prepr.	Preprint
RDF	Resource Description Framework

RVK	Regensburger Verbundklassifikation
s., s.a.	siehe, siehe auch
s.o., s.u.	siehe oben, siehe unten
SBIG	Subject Based Information Gateway
SGML	Standard Generalized Markup Language
SIC	Standard Industrial Classification
sl.	slide(s)
SMART	System for Manipulating And Retrieving Text
SOSIG	Social Science Information Gateway
Tab.	Tabelle(n)
TFIDF	Termfrequenz * Inverse Dokumentenfrequenz
TU	Technische Universität
u.a.	unter anderem, und andere
u.a.m.	und anderes mehr
u.dgl.	und dergleichen
U.S.	United States
UB	Universitätsbibliothek
UDC, UDK	Universal Decimal Classification, Universale Dezimalklassifikation
UKOLN	UK Office for Library and Information Networking (Univ. Bath)
Univ.	Universität
URL	Uniform Resource Locator
USPC	United States Patent Classification
USPTO	U.S. Patent and Trademark Office (Washington, DC)
usf., usw.	und so fort, und so weiter
Verl.	Verlag
vgl.	vergleiche
vs.	versus
WAIS	Wide Area Information Server
WIPO	World Intellectual Property Organization (Genf)
WWLib	Wolverhampton Web Library
WWLib-TNG	WWLib – The Next Generation
WWLib-TOS	WWLib – The Original Software
WWW	World Wide Web
XHTML	Extensible Hypertext Markup Language
z.B.	zum Beispiel
z.T.	zum Teil
ZDF	Zweites Deutsches Fernsehen
zit. n.	zitiert nach

1 Einleitung

1.1 Thematischer Hintergrund

Die rapide Zunahme von elektronischen Dokumenten (insbesondere von Web-Ressourcen) hat zu einem Wiederaufleben des Interesses an Klassifikationen bzw. klassifikationsähnlichen Schemata für die Erschliessung dieser Ressourcen und ihre Präsentation in einer hierarchischen thematischen Browsing-Struktur geführt. Dies manifestierte sich am deutlichsten durch die Entstehung von Web-Katalogen ("Web directories", "subject trees") wie *Yahoo!*,[1] die Ende der 1990er Jahre weit häufiger benutzt wurden als die damals populärsten stichwortorientierten Suchmaschinen (Koch 1998b, 326). Dies mag sich durch die Erfolge einer neuen Generation von Suchmaschinen wie z.B. *Google*[2] in den letzten Jahren etwas verschoben haben, doch üben die hierarchisch strukturierten Web-Kataloge (*Yahoo!, LookSmart, Seeq, DMOZ* usw.) bis heute eine starke Präsenz aus – und auch Dienste wie *Google* bieten inzwischen neben der stichwortorientierten Suche auch hierarchisch strukturierte Web-Verzeichnisse an.

Dieser Erfolg kommt nicht von ungefähr. In der bibliothekarischen Welt sind die Vorzüge einer klassifikatorischen Erschliessung seit jeher wohlbekannt; nun wurden sie auch von Informatikern und Web-Entwicklern entdeckt: "The most successful paradigm for organizing this mass of information, making it comprehensible to people, is by categorizing the different documents according to their topic, where topics are organized in a hierarchy of increasing specificity" (Koller & Sahami 1997, 170).

Bei der Entwicklung der klassifikatorischen Strukturen für diese Web-Verzeichnisse spielten etablierte Klassifikationssysteme aus dem Bibliotheksbereich jedoch keine wesentliche Rolle. Dies lag zum einen daran, dass die international bedeutendsten Systeme (DDC und LCC) erst relativ spät in maschinenlesbare Form konvertiert worden waren (Vizine-Goetz 1999b), zum anderen aber wohl auch daran, dass seit Entstehen des WWW dessen treibende Kräfte die Tendenz aufwiesen, das Rad neu zu erfinden ("often badly") und dass in der "Internetkultur" der Glaubensgrundsatz galt, dass Bibliothekare zu diesem neuen Zeitalter nichts beizutragen hätten (Bates 2002).

Während *Yahoo!* stellvertretend für den Erfolg der klassifizierten Web-Kataloge steht, verkörpert dieser Dienst gleichzeitig aber auch die "Krise dieser Tätigkeit auf globalem Niveau" (Koch 1998b), da Auswahl und klassifikatorische Erschliessung der jeweils nachgewiesenen Ressourcen bis heute manuell bzw. intellektuell erfolgen. Dies hat zwar im Vergleich zu Suchmaschinen deutliche Vorteile hinsichtlich der Qualität

[1] http://www.yahoo.com [16.05.2004]
[2] http://www.google.com [16.05.2004]

des Nachgewiesenen, aber auch drastische Nachteile bezüglich der Quantität zur Folge: "Wer durchsucht schon gerne (ohne Not) eine Untermenge einer Datenbank, wenn er alles haben kann?" (Lepsky 1998, 336).

Vor diesem Hintergrund begann man sich in den 1990er Jahren mit Ansätzen zur *Automatisierung* der klassifikatorischen Erschliessung von Web-Dokumenten zu beschäftigen, da eine Klassifizierung von elektronischen Ressourcen auf breiter Basis nur auf diese Weise vorstellbar ist: "The combination of automation and classification has the potential to provide an accurate, intuitive, comprehensive classified search engine" (Jenkins et al. 1997, 2). Da diese Projekte zum Teil von bibliothekarischer Seite ausgingen, spielten etablierte Klassifikationssysteme wie DDC und UDK nun wieder eine Rolle, zumal sich diese inzwischen auch in verschiedenen manuell erschlossenen virtuellen Bibliotheken bzw. SBIGs[3] bewährt hatten (CyberStacks, BUBL, SOSIG, OMNI usw.)

Automatisches Klassifizieren wird typischerweise als Vergleich von Repräsentationen von Dokumenten mit solchen von Klassen mittels eines Ähnlichkeitsmasses durchgeführt. Dafür steht eine Reihe verschiedener methodischer Ansätze und Verfahren zur Auswahl, auf die im Rahmen dieser Arbeit näher eingegangen werden soll.

1.2 Terminologie und Definitionen

Die wichtigen Definitionen und Begriffsabgrenzungen zum automatischen Klassifizieren werden in Abschnitt 2.1.1 gegeben. Vorausschickend seien jedoch an dieser Stelle zwei Aspekte festgehalten:

- "Automatisches Klassifizieren" wird im Rahmen dieser Arbeit im Sinne des Klassifizierens von *Text*dokumenten verstanden – das Klassifizieren z.B. multimedialer Dokumente wird hier nicht betrachtet;
- "Automatisches Klassifizieren" soll als Verfahren zur Einordnung von Dokumenten in *vorgegebene* Klassen eines Klassifikationssystems verstanden werden und *nicht*, wie gelegentlich in der Literatur praktiziert, auch das *Clustern* von Dokumenten (d.h. das Auffinden einer Klassenstruktur in der Dokumentenmenge) inkludieren.

Bei der Verwendung der bibliothekarischen Fachterminologie wird der mehrdeutige Terminus "Klassifikation" vermieden. Stattdessen wird von *Klassifizieren* bzw. *Klassifizierung* gesprochen, wenn es um die Tätigkeit bzw. den Prozess der Zuordnung zu Klassen geht, sowie von *Klassifikationssystem*, wenn die strukturierte Darstellung von Klassen und der zwischen ihnen bestehenden Begriffsbeziehungen gemeint ist. In der vielfältigen Literatur zum Thema "automatisches Klassifizieren" werden die Be-

[3] "Subject based information gateways", d.h. "fachbezogene Qualitätsdienste" (Koch 1997a) oder "qualitätskontrollierte Fachinformationsdienste" (Koch 1998a).

griffe "classification", "categorization" und "taxonomy" sehr häufig synonym verwendet. Die mit einer solchen Gleichsetzung verbundene Problematik kann im Rahmen dieser Arbeit jedoch nicht weiter analysiert werden. Mit "classification" wird in der Literatur manchmal auch "Klasse", "Kategorie" oder "Notation" gemeint; hier erfolgt jedenfalls eine Unterscheidung zwischen den ersten beiden Bedeutungen einerseits und der dritten andererseits.

Der Begriff *elektronische Dokumente* meint zwar in erster Linie Web-Ressourcen, inkludiert aber grundsätzlich alle Formen elektronischer (Volltext-)Dokumente. Da die Arbeit mit einem bibliothekarisch-informationswissenschaftlichen Hintergrund bzw. Interesse verfasst ist, sollen hier aber auch bibliographische Datensätze, wie sie in Online-Katalogen vorkommen, als elektronische Dokumente gelten dürfen.

Im Gegensatz zu der in Literatur und Alltagssprache vielfach zu beobachtenden Praxis wird in dieser Arbeit der Begriff *Internet* nicht (fälschlich) als Synonym für *WWW* bzw. *Web* gebraucht.

1.3 Zielsetzung

Die vorliegende Arbeit zielt darauf ab, den Themenkomplex "automatisches Klassifizieren" näher zu durchleuchten und das Potential dieses methodischen Ansatzes aus bibliothekarischer und informationswissenschaftlicher Sicht auszuloten. Sind die Techniken des automatischen Klassifizierens heute bereits so weit, dass damit grosse Mengen elektronischer Dokumente (Web, Bibliotheken, Dokumentation) zufriedenstellend erschlossen werden können? Wie ist der gegenwärtige Stand einschlägiger Anwendungen?

Das *primäre* Ziel der der Arbeit besteht in der Erarbeitung einer *informierten und aktuellen Sichtweise* zu diesem Themenkomplex. Dies soll durch die Beantwortung der folgenden *Fragen* geschehen:

a) Was wird *heute* in methodischer Hinsicht unter automatischem Klassifizieren verstanden? Welche *Techniken* kann bzw. sollte man einsetzten, wenn elektronische Dokumente ohne (oder fast ohne) menschliche Intervention nach einem Klassifikationssystem erschlossen werden sollen?

b) Was wurde in praktischer Hinsicht bis heute erreicht? Was geschah "wirklich" (d.h. im Detail) im Rahmen der wenigen "grossen", bekannten und jeweils durch eine Vielzahl von Publikationen – aber keine detaillierten Gesamtdarstellungen – dokumentierten Anwendungsprojekte?

c) Welche sonstigen Anwendungen und Projekte sind aus bibliothekarisch-informationswissenschaftlicher Sicht von Interesse? Gibt es Ansätze zur automatischen

Klassifizierung von Bibliotheksbeständen, in der Dokumentation etc.? Wie wird bzw. wurde dabei methodisch vorgegangen und welche Resultate wurden erzielt?

d) Wie können die jeweils erhobenen Befunde bewertet werden?

Zur Beantwortung dieser Fragen soll die Arbeit folgende *Ergebnisse* liefern:

- Eine kompakte und aktuelle Übersicht über die Methodik des automatischen Klassifizierens;
- Eine ausführliche Darstellung und Analyse der vier grossen und bekannten Projekte zur automatischen Klassifizierung von elektronischen (Web-)Dokumenten;
- Eine Darstellung und Analyse weiterer Anwendungen und Projekte auf Gebieten wie Bibliothekswesen, Patentdokumentation, Mediendokumentation, WWW usw.;
- Eine kritische Diskussion und Beurteilung dieser Befunde bzw. sonstiger relevanter Aspekte.

Die *zweite* Zielsetzung der Arbeit ist eine *didaktische*, zumal es ihre deklarierte Absicht ist, das Thema, die Methodik, die praktischen Anwendungen und Vorgangsweisen für Leser, die zwar über bibliothekarisch-informationswissenschaftliche Vorkenntnisse, aber kein detailliertes Vorwissen aus dem Bereich des automatischen Klassifizierens verfügen (wie auch der Autor selbst), *verständlich* darzustellen. Vorausgesetzt werden dabei allerdings sehr wohl Grundbegriffe aus den Bereichen inhaltliche Erschliessung, Information Retrieval, WWW und Internet.[4]

1.4 Thematische Eingrenzung

Obwohl es die Absicht der Studie ist, einen umfassenden Überblick über das Themengebiet zu bieten, sollen einige Eingrenzungen nicht unerwähnt bleiben:

- In *inhaltlicher* Hinsicht geht es in erster Linie um die Anwendbarkeit der besprochenen Verfahren für die inhaltliche Erschliessung elektronischer Dokumente; die sehr reichhaltige und umfangreiche Informatik-Literatur zum automatischen Klassifizieren, die sich vor allem mit methodischen Detailfragen beschäftigt, wurde dabei nur unter dem erstgenannten Gesichtspunkt genutzt.
- Im Hinblick auf *Umfang* und *Vollständigkeit* wurde zwar versucht, die relevanten Projekte und Anwendungen möglichst gut zu erfassen, doch kann nicht der Anspruch erhoben werden, dass dies lückenlos gelungen ist.

[4] So etwa Begriffe und Themen wie Thesaurus, Klassifikation (DDC, LCC, UDK, Facettenklassifikation); Vektorraummodell, Termgewichtung, Relevance Ranking, Precision und Recall; Internet-Dienste, Metadaten, HTML, URL, Harvesting, Suchmaschinen.

- In *sprachlicher* Hinsicht basiert die Arbeit praktisch ausschliesslich auf deutsch- und englischsprachiger Literatur, obwohl vermutlich auch in anderen Sprachen interessante Publikationen zum Thema vorliegen.
- In *zeitlicher* Hinsicht reicht die Betrachtung bis etwa in die erste Hälfte der 1990er Jahre zurück, da damals das Interesse an derartigen Vorhaben aufzuleben begann. Als "Startzeitpunkt" kann die Studie von Larson (1992) gelten, in der erstmals Ansätze zur automatischen Vergabe von LCC-Notationen auf der Basis von Titel- und LCSH-Information in MARC-Datensätzen experimentell untersucht wurden (vgl. Abschnitt 7.1.1). Davor war es den Autoren und Forschungsgruppen primär um das Clustern von Dokumenten *per se* oder um den Vergleich solcher Cluster-Ergebnisse mit einer bibliotheksseitig vorgenommenen Klassifizierung gegangen (ibid., 131).

1.5 Methodische Vorgangsweise

Der grösste Teil dieser Arbeit basiert auf der Auswertung und Analyse publizierter wie auch grauer Literatur. Diese wurde durch ausführliche und systematische Recherchen auf folgende Art gefunden:

- Verfolgung der Literaturhinweise in den bereits bekannten und den im Zuge der Recherchen aufgefundenen Arbeiten;
- Recherche in bibliographischen Datenbanken auf CD-ROM[5] bzw. via WWW;[6]
- Web-Recherche mittels Suchmaschinen;[7]
- Systematische Überprüfung der Webseiten aller bekannten einschlägigen Projekte.

Daneben wurden auch – soweit via WWW zugänglich – verschiedene Prototypen und Anwendungen (*EELS*, *WWLib*, *GERHARD* und weitere) auf ihre Funktionalitäten hin untersucht. Nur im Fall des Projektes *GERHARD* (vgl. Kapitel 5) wurden auch einige Informationen telefonisch eingeholt.

Die bisher vorliegenden Übersichten (Davis et al. 2003; Gietz 2001, Hoffmann 2002, McKiernan 1996, Robbins 1999, Srishaila 2001, Tóth 2002, Woodward 1996) veranschaulichen die hier analysierten Projekte und Anwendungen nur unzureichend oder oberflächlich. Die Darstellung basiert daher primär *nicht* auf diesen Quellen, sondern auf der jeweils veröffentlichten Projektliteratur.

[5] LISA, Information Science Abstracts Plus, INSPEC.

[6] ERIC, DOBI, Current Cites.

[7] hauptsächlich *Google* (http://www.google.com/)

1.6 Gliederung der Arbeit

Die Studie beginnt mit einem Überblick über den gegenwärtigen Stand der Methodik des automatischen Klassifizierens (Kapitel 2). Danach werden die bisher durchgeführten grossen Projekte zur automatischen Erschliessung von Web-Dokumenten an den Universitäten Lund, Wolverhampton und Oldenburg sowie bei OCLC im Detail analysiert (Kapitel 3 bis 6). Kapitel 7 beschäftigt sich mit weiteren Anwendungen und Projekten im Umfeld von Bibliothek, Dokumentation und Informationspraxis. In Kapitel 8 wird schliesslich der Versuch unternommen, diese Vorhaben bewertend zusammenzufassen und einige weitere Aspekte des Einsatzes automatischer Klassifizierungsverfahren zu diskutieren.

2 Zur Methodik des automatischen Klassifizierens

In diesem Kapitel wird der gegenwärtige Stand der Methodik des automatischen Klassifizierens von Textdokumenten in einem knappen, aber dennoch umfassenden Überblick dargestellt. Diese Betrachtung verfolgt zwei Zielsetzungen: Zum einen soll die Thematik, die primär von unserer Nachbardisziplin *Informatik* besetzt ist, für Interessenten aus dem Bereich der Informationswissenschaft und -praxis *verständlich* aufbereitet werden. Die Darstellung vermeidet daher nahezu jegliche formalisierte Ausdruckweise, wie sie in der einschlägigen Informatik-Literatur in der Regel vorzufinden ist. Zum anderen soll die systematische Behandlung der Methodik eine Informationsgrundlage für die Einordnung und Interpretation der in den folgenden Kapiteln referierten praktischen Anwendungen und Projekte bilden.

Die folgenden Ausführungen stützen sich, was Aufbau und inhaltliche Abfolge betrifft, vor allem auf den umfangreichen Übersichtsartikel von Sebastiani (2002b) und verdanken diesem und einer Reihe weiterer Beiträge[1] wesentliche Informationen.

2.1 Automatisches Klassifizieren von Textdokumenten

2.1.1 Begriffsumfang

Textklassifizierung[2] – im Englischen als "text categorization", "text classification" oder (seltener) "topic spotting" bezeichnet – wird definiert als "the activity of labeling natural language texts with thematic categories from a predefined set" (Sebastiani 2002b, 1). Die Beschäftigung mit der *Automatisierung* dieses Prozesses reicht bereits mehrere Jahrzehnte zurück. Ursprünglich galt die Klassifizierungsaufgabe dabei als eine durch ein IR-System zu lösende, wobei ein neu zu klassifizierendes Dokument (repräsentiert durch einen Vektor von Termgewichten) als "Anfrage" an das System gerichtet wurde, das einen Vergleich mit den in einer Datenbank gespeicherten Klassenvektoren (Zentroiden) der als Cluster aufgefassten Klassen durchführte und eine nach Ähnlichkeit ranggeordnete Liste der Klassen erstellte (Salton & McGill 1987, 145f.) In den 1980er Jahren dominierte ein Ansatz, der auf der Verwendung von Expertensystemen basierte, die über manuell definierte Regeln die Zuteilung der zu klassifizierenden Dokumente zu den jeweiligen Klassen vornahmen. Seit den 1990er Jahren hat sich, zunächst in der Forschung und inzwischen auch bereits im Umfeld kommerzieller Software, der Einsatz

[1] Brückner (2001), Delphi Group (2002), Goller et al. (2000), Hoffmann (2002), Klinkenberg (1998), Ravid (2002), Recommind (o.J.), Renz (2001), Sebastiani (2002a; 2004), Yang (1999) und Yang & Liu (1999).

[2] In der deutschsprachigen Literatur meist "...klassifikation".

maschineller Lernverfahren[3] durchgesetzt. Ein darauf basierendes automatisches Klassifizierungssystem besteht stets aus zwei Komponenten:

- Die Komponente zum *Wissenserwerb* wird in der *Trainingsphase* eingesetzt, in welcher auf der Grundlage einer Menge bereits intellektuell klassifizierter *Trainingsdokumente* die Charakteristika der Klassen gelernt und Klassenprofile erstellt werden.
- Die eigentliche Komponente zum *Klassifizieren* ("Klassifikator") wird dann in der darauffolgenden *Klassifizierungsphase* eingesetzt, in der neue, d.h. noch nicht klassifizierte Dokumente hinsichtlich ihrer Charakteristika analysiert und durch einen Vergleich mit den Klassenprofilen den passenden Klassen (oder auch nur einer Klasse) zugeordnet werden.

Im Gegensatz hiezu wird der Begriff "automatisches Klassifizieren" manchmal auch für die automatische Identifizierung der Klassenstruktur einer Dokumentenkollektion bzw. die Einordnung der Dokumente in eine solche neu gefundene Struktur verwendet. Für Verfahren dieser Art, die üblicherweise unter "text clustering" – im Deutschen *Clustern* oder *Clustering* – subsumiert werden, soll, wie schon erwähnt, in dieser Arbeit der Begriff jedoch nicht gelten; hier soll es ausschliesslich um die automatische Zuordnung zu Klassen bereits bestehender Klassifikationssysteme gehen.

In den "grossen" Projekten zum automatischen Klassifizieren, die in den folgenden vier Kapiteln analysiert werden und die sämtlich aus den 1990er Jahren stammen, wurden die für den Klassifizierungsprozess benötigten Klassenprofile – mit der Ausnahme weniger neuerer Projektphasen – nicht mittels lernender Verfahren, sondern *manuell* erstellt. Dies geschah zumeist durch mehr oder weniger raffinierte Techniken der Vokabularanreicherung, auf die unten noch im Detail eingegangen wird. Begreift man diese Vorgangsweise als Prozess, so ist dabei die semantische Bedeutung der Klassen ständig präsent und von entscheidender Rolle.

Im Gegensatz dazu gehen maschinelle Lernverfahren davon aus, dass die Klassen nur symbolische "labels" sind[4] und weitere Informationen über ihre Bedeutung ausschliesslich aus dem Text der bereits klassifizierten Trainingsdokumente – dem *endogenen* Wissen – gewonnen werden können. In der Praxis wird jedoch bei Vorliegen von *exogenem* Wissen – Informationen aus anderen Quellen (Klassenbenennungen in Tafeln, Register, Metadaten usw.) – durchaus versucht, mithilfe heuristischer Verfahren auch dieses zu nutzen.

Das Ziel des automatischen Klassifizierens ist die Approximation einer unbekannten Zielfunktion ("target function"), die beschreibt, wie die Dokumente klassifiziert werden sollten. Dies geschieht mittels einer weiteren Funktion, die üblicherweise als

[3] Maschinelles Lernen ist "... the study of computer algorithms that automatically improve performance through 'experience'" (Ravid 2002, sl.12).

[4] Daher wird in der einschlägigen Literatur praktisch *nicht* zwischen "Klassifikationssystem", "Taxonomie" und "Kategorienschema" unterschieden. Anstelle von "Klassen" wird häufig auch von "Kategorien" oder "labels" gesprochen.

Klassifikator ("classifier") bezeichnet wird,[5] und zwar so, dass diese beiden Modelle so gut wie möglich übereinstimmen. Masse für diese Güte ("effectiveness") werden in Abschnitt 2.6.1 behandelt.

2.1.2 Einfach- und Mehrfachklassifizierung

Wenn jedem Dokument nur genau *eine* Klasse zugewiesen werden soll, spricht man von "single-label categorization" (Einfachklassifizierung). Ein Spezialfall davon ist "binary classification" – jedes Dokument wird in eine von zwei Klassen einsortiert. "Multilabel categorization" bezeichnet dagegen den Fall, in dem die Klassenzahl von null bis zum Maximum variieren kann ("overlapping categories"). Der binäre bzw. einfache Fall ist allgemeiner als der multiple, da ein Verfahren für eine binäre Klassifizierung auch für die multiple verwendet werden kann, indem die Klassifizierungsaufgabe in mehrere unabhängige Aufgaben zerlegt wird (die stochastische Unabhängigkeit der Klassen von einander wird dabei unterstellt). Daher konzentriert sich die einschlägige Literatur meist auf die Betrachtung des binären bzw. einfachen Falles.

2.1.3 Klassen- vs. Dokumentenzentrierung

Bei der Verwendung eines Textklassifikators bestehen zwei Möglichkeiten:

- Suche nach allen Klassen, denen ein Dokument zugeordnet werden soll (Dokumentenzentrierung);
- Suche nach allen Dokumenten, die in eine Klasse sortiert werden sollen (Klassenzentrierung).

Der erstgenannte Fall ist der häufigere; der zweite ist z.B. gegeben, wenn die Klassenstruktur um eine neue Klasse erweitert werden soll und die bereits klassifizierten Dokumente neuerlich zu prüfen sind. Einzelne Methoden zur Konstruktion des Klassifikators weisen eine Tendenz zum einen oder anderen Fall auf. In der dieser Arbeit zugrundeliegenden Literatur werden diese beiden Sichtweisen aber häufig vermischt.

2.1.4 "Harte" vs. rangordnende Klassifizierung

Während *vollautomatische* Systeme eine "harte" Klassifizierung benötigen und deshalb eine Zugehörigkeitsentscheidung für jede Dokument-Klasse-Kombination treffen müssen, liefern andere lediglich eine ranggeordnete Liste der Klassen in bezug auf ihr Zutreffen je Dokument (bzw. ein Ranking der Dokumente hinsichtlich ihres Zutreffens je Klasse), ohne eine solche "harte" Entscheidung vorzunehmen. Letzteres ist vor allem dann sinnvoll, wenn der Output als Entscheidungsgrundlage für eine intellektuelle Klas-

[5] "Klassifikator" steht für das klassifizierende Verfahren: "A classifier inputs a document and outputs a class" (Chakrabarti et al. 1998, 167).

sifizierung dienen soll, da sich in diesem Fall der menschliche Auswahlprozess auf die *k* bestgereihten Klassen stützen kann, ohne das ganze Klassifikationssystem betrachten zu müssen. Man spricht in diesem Fall auch von *semi-automatischer* Klassifizierung. Abweichend davon wird dieser Begriff auch für jene Fälle verwendet, in denen der Klassifikator nicht automatisch, sondern manuell erstellt wird (z.B. Attardi et al. 1999, 2).

2.2 Hauptanwendungen der automatischen Textklassifizierung

Sebastiani (2002b) unterscheidet fünf hauptsächliche Anwendungsgebiete der automatischen Klassifizierung von Texten:

Automatisches Indexieren für Boolesche Textretrievalsysteme: Zuordnung von Dokumenten zu Termen eines kontrollierten Vokabulars (z.B. Thesaurus), wobei diese Terme (Deskriptoren) als *Kategorien* gelten, sodass automatisches Indexieren als eine Instanz des automatischen Klassifizierens gesehen werden kann.[6] Anwendungen dieser Art bildeten den Schwerpunkt der frühen Ansätze zum automatischen Klassifizieren (1960er und 1970er Jahre), finden sich aber durchaus auch noch heute (z.B. in der Mediendokumentation, vgl. Abschnitt 7.3.1). Eine moderne Variante dieser Anwendung stellt z.B. das automatische Generieren von thematischen Metadaten anhand eines Thesaurus dar.

Dokumentenorganisation: Die Indexierung mit kontrolliertem Vokabular ist ein Spezialfall der Ordnung von Dokumenten. Viele andere Aktivitäten zum Zweck der Dokumentenorganisation können mit mittels automatischer Textklassifizierung erfolgen. Beispiele sind etwa die automatische Zuordnung von Kleinanzeigen zu thematischen Rubriken in Zeitungen bzw. Magazinen; die Sortierung einlangender Patentanmeldungen zum Zweck der Weiterbearbeitung durch fachliche Teams (vgl. Abschnitt 7.2) und die automatische Kategorisierung von Zeitungsartikeln (vgl. 7.3).

Thematisches Textfiltern: Unter "topic filtering" wird das Klassifizieren eines Stroms einlangender Dokumente verstanden, z.B. der von Nachrichtenagenturen an Zeitungsredaktionen ausgesandten Meldungen ("newsfeed"). Das Ausfiltern der nicht erwünschten Teile dieses Nachrichtenstroms kann als "single-label" Textklassifizierung gesehen werden, bei der die einlangenden Texte in die beiden disjunkten Klassen "relevant" und "irrelevant" sortiert werden. Dabei können die relevanten Dokumente noch weiter in thematische Klassen gruppiert werden. Ein anderes, aktuelles Beispiel ist das Filtern von elektronischer Post (erwünschte vs. unerwünschte Nachrichten). Wenn bei solchen An-

[6] Umgekehrt kann auch Klassifizieren als eine Form des Indexierens gesehen werden (vgl. DIN 31623, Teil 1, 3.5; im Kontext automatischer Verfahren z.B. Nohr 2003, 29–32).

wendungen Benutzerprofile verwendet und aufgrund von Benutzerfeedback verändert werden, spricht man von *adaptivem Filtern* ("adaptive filtering").[7]

Disambiguierung der Bedeutung von Homonymen bzw. Polysemen: Wenn die Kontexte der auftretenden Wörter als Dokumente und die Wortbedeutungen als Klassen betrachtet werden, so kann die Disambiguierung als Aufgabe einer Einfachklassifizierung (dokumentenzentriert) gesehen werden. Zu diesem Anwendungsbereich gehören etwa auch die kontextsensitive Korrektur von Schreibfehlern, die Markierung/Extraktion von Phrasen ("part of speech tagging") oder die Wortwahl bei der automatischen Übersetzung.

Hierarchisches Klassifizieren von elektronischen (v.a. Web-)Dokumenten: Hiezu zählen die in dieser Arbeit hauptsächlich betrachteten Anwendungen und Projekte. Die hierarchische Struktur des Klassifikationssystems wird dabei oft so berücksichtigt, dass das gesamte Klassifizierungsproblem in eine Reihe kleinerer Klassifizierungsprobleme aufgelöst wird, die jeweils einer Verzweigungsentscheidung an einem Knoten des Systems entsprechen (vgl. z.B. Ceci & Malerba 2003; Dumais & Chen 2000; Frank & Paynter 2004; Greiner et al. 1997; Koller & Sahami 1997). Dies gilt als wesentlich effizienter als das mitunter geübte Ignorieren der hierarchischen Struktur (d.h. die Annahme einer flachen Kategorienstruktur). In jüngerer Vergangenheit hat sich in mehreren Studien auch herausgestellt, dass die Einbeziehung der hypertextuellen Struktur von Web-Dokumenten beim Klassifizieren von Vorteil sein kann (vgl. Abschnitt 2.6.3). Das Klassifizieren von Web-Dokumenten kann mit dem thematischen Textfiltern zusammenfallen, wenn der Klassifikator bereits im Rahmen des Harvesting der Dokumente aus dem Web eingesetzt wird (vgl. Abschnitt 3.2.7).

2.3 Maschinelle Lernverfahren

Der bis in die 1990er Jahre insbesondere bei operational eingesetzten Systemen verbreitete Ansatz des automatischen Klassifizierens sah die Verwendung von Expertensystemen mit der Fähigkeit zur Lösung von Klassifizierungsaufgaben vor. Ein solches Expertensystem beinhaltete eine grosse Zahl von manuell definierten logischen Regeln (d.h. eine oft komplexe Regeldefinition pro Klasse), die darüber entschieden, ob ein Dokument einer Klasse zugeordnet werden sollte oder nicht.[8] Wenn das Dokument zumindest eine der Klauseln der betreffenden Regel erfüllte, so wurde es der entsprechenden Klasse zugeteilt. Diese Vorgangsweise war jedoch sehr ressourcenintensiv, da die Er-

[7] Der in der praktischen Dokumentation und Information seit langem bekannte Dienst "selective dissemination of information" bzw. "current awareness" stellt ebenfalls eine Form des Dokumentfilterns dar.

[8] z.B. für die Klasse "Weizen": **if** ((wheat & farm) **or** (wheat & commodity) **or** (bushels & export) **or** (wheat & tonnes) **or** (wheat & winter &¬ soft)) **then** WHEAT **else** ¬ WHEAT (nach Sebastiani 2002b, 8).

stellung der Regeln die Zusammenarbeit hochqualifizierter Vertreter aus Knowledge Engineering und Fachdisziplin erforderte. Andererseits wurden (zumindest im Laborexperiment) mit dieser Technik mitunter hervorragende Ergebnisse erzielt, die sogar die besten mit den modernen lernenden Klassifikatoren erreichten Resultate überflügelten.

Im Rahmen des heute dominierenden, auf dem maschinellen Lernen basierenden Ansatzes erstellt ein induktiver Prozess ("learner") automatisch einen Klassifikator für jede Klasse, indem er die Eigenschaften einer Menge von Dokumenten analysiert, die zuvor durch menschliche Experten in die betreffende Klasse eingeordnet wurden. Auf dieser Grundlage werden jene Eigenschaften ausfindig gemacht, die ein neues, noch nicht klassifiziertes Dokument ("unseen document") aufweisen sollte, um der betreffenden Klasse zugeordnet zu werden. Man spricht in diesem Fall von *überwachtem* Lernen ("supervised learning"), da der Lernprozess durch das in den Klassen und in den zugehörigen Trainingsdokumenten enthaltene Wissen gleichsam überwacht wird.[9] Die bereits klassifizierten Trainingsdokumente stellen demzufolge die wichtigste Ressource für diesen Ansatz dar.

2.3.1 Trainings-, Test- und Validierungsdokumente

Die Menge der Trainingsdokumente ("training set"), die einem lernenden Klassifikator zugrundegelegt wird, muss Beispiele, d.h. bereits intellektuell klassifizierte Dokumente, für jede Klasse enthalten. Ein zu einer bestimmten Klasse zählendes Dokument gilt als *positives* Beispiel, ein nicht zu dieser Klasse zählendes als *negatives* Beispiel; eine Reihe von Algorithmen kann von beiden Typen Gebrauch machen. Im Fall von Mehrfachklassifizierung kann hier ein Problem auftreten, da die zu einer bestimmten Klasse zählenden positiven Beispieldokumente zu einem gewissen Grad von negativen Beispielen durchsetzt sind, da einige von ihnen auch zu anderen Klassen gehören (vgl. z.B. Koster et al. 2003).

Wenn – wie zumeist – der Wunsch besteht, die Güte der durch das automatische Verfahren erzielten Ergebnisse zu evaluieren, wird *vor* der Erstellung des Klassifikators die Trainingsmenge in zwei (nicht notwendigerweise gleiche) Teile geteilt:

- die *Trainingsdokumente*; auf der Basis dieser Dokumente wird der Klassifikator gebildet; soferne es in der Trainingsphase notwendig ist, verschiedene Versionen oder Parameter des Klassifikators zu testen, wird zuvor den Trainingsdokumenten eine Menge von *Validierungsdokumenten* entnommen, anhand derer diese Tests durchgeführt werden;
- die *Testdokumente*; diese werden für den Test auf Güte herangezogen, indem jedes Dokument automatisch klassifiziert und das Resultat mit der intellektuellen Klassifizierung verglichen wird (zu den dabei verwendeten Massen s.u.)

Meist ist die Trainingskollektion deutlich grösser als die Testkollektion. Die letztere darf keinesfalls bei der Erstellung des Klassifikators mitwirken, da dies den an-

[9] Ein Fall von "unsupervised learning" wäre dagegen das oben erwähnte Clustering.

schliessenden Test verfälschen würde. Dies gilt sinngemäss auch für eine allfällig verwendete Validierungskollektion. Nach der Evaluierung wird üblicherweise der Klassifikator neuerlich – und diesmal auf der Basis des gesamten Korpus – trainiert, um seine Leistung mittels Vergrösserung der Trainingsmenge noch zu verbessern.

2.3.2 Information-Retrieval-Techniken und Textklassifizierung

Bei der automatischen Textklassifizierung werden mehrere Techniken aus dem klassischen IR eingesetzt; dies geschieht in den folgenden Phasen:

- die Aufbereitung der Dokumente des Trainings- wie auch des Testkorpus (bzw. in der Folge der im operationalen Betrieb zu klassifizierenden Dokumente) erfolgt durch eine IR-typische Indexierung;
- bei der induktiven Konstruktion des Klassifikators werden häufig IR-typische Techniken eingesetzt (z.B. "document-request matching", "query reformulation");
- bei der Evaluierung der Klassifizierungsgüte werden in der Regel aus dem IR bekannte Masse verwendet.

2.4 Dokumentenindexierung und Merkmalsreduktion

2.4.1 Indexierung der Dokumente

Damit ein Klassifikator bzw. ein Algorithmus, der einen Klassifikator erstellt, die zu verwendenden Texte interpretieren kann, wird ein Indexierungsverfahren benötigt, das eine kompakte Repräsentation dieser Texte zum Ergebnis hat. Üblicherweise geschieht dies durch die Bildung eines *Vektors von Termgewichten*, der ausdrücken soll, zu welchem Ausmass jeder im Dokument auftretende Term – oft *Attribut* oder *Merkmal* ("feature") genannt – zur Bedeutung des betreffenden Trainings- bzw. Testdokuments beiträgt. Dabei gibt es verschiedene Möglichkeiten für die Definition eines Attributs sowie für die Berechnung der Termgewichte.

Am häufigsten werden einfach die einzelnen Wörter aus dem Text als die Attribute definiert; man spricht in diesem Fall von einem "bag-of-words"-Ansatz.[10] Die (eigentlich naheliegende) Verwendung von Phrasen anstelle einzelner Wörter hat bislang nicht zu wesentlich besseren Ergebnissen geführt. Am besten scheint dabei noch die Kombination von syntaktischen und statistischen Phrasen (d.h. grammatikalisch gebildeten bzw. auf Basis ihres gemeinsamen Vorkommens ermittelten Wortgruppen) abzuschneiden.

Die Gewichte werden meist auf einen Wertebereich zwischen "0" und "1" normiert. Je nach Verfahren kommen binäre Gewichte (1 = Term tritt auf; 0 = Term tritt nicht auf) oder nichtbinäre Gewichte zur Verwendung. Im Prinzip kann jedes Verfahren

[10] Für das Zerlegen eines Textes in einzelne Wörter wird häufig auch der Begriff "Tokenisierung" verwendet; vgl. z.B. Dörre et al. (2001, 435).

angewandt werden, welches zur Darstellung der Dokumente als Vektoren aus gewichteten Termen führt. Am häufigsten ist dies die bekannte *TFIDF*-Funktion, die die Häufigkeit eines Terms im Dokument (je häufiger, desto bedeutender für den Inhalt) zur Zahl der Dokumente, in denen der Term auftritt (je öfter, desto geringere Trennschärfe), in Relation setzt (vgl. Salton & McGill 1987, 68f.)

Vor der Indexierung wird meist eine *Textnormalisierung* durchgeführt, bei der alle unerwünschten Zeichen (z.B. HTML-Kodierung) bzw. Terme entfernt werden. Dabei gelangen häufig auch linguistische Verfahren zur Anwendung, insbesondere die Eliminierung von Stoppwörtern (sehr häufige und/oder sehr seltene Wörter; Wörter mit geringer Buchstabenzahl) und die Lemmatisierung bzw. Stammformenbildung ("stemming"). Bei der Behandlung deutschsprachiger Texte sollten hiezu auch noch Techniken wie Kompositazerlegung, Derivation (Zusammenfassen von verschiedenen Wortklassen oder Derivaten, wie z.B. Adjektiven, Substantiven und Verben, mit derselben Grundform) und Bindestrichergänzung treten.

Zur Repräsentation der Dokumente wird entweder der gesamte Text (Volltext) oder nur ein Teil des Textes herangezogen. Im letzteren Fall können dies etwa bestimmte Textabschnitte wie Titel, HTML-Headings, Abstract usw. sein, die manchmal auch noch mit unterschiedlichen Gewichtungsschemata belegt werden, sodass etwa Wörter aus dem Titel oder Überschriften mehr Gewicht erhalten als solche aus dem übrigen Text. Je strukturierter die zu klassifizierenden Dokumente sind, desto eher ist eine derartige Vorgangsweise realisierbar.

2.4.2 Dimensionsreduktion

Da die bei Verwendung einer Vielzahl von Termen resultierende hohe Dimensionalität des Merkmalsraums für viele der zur Erstellung eines Klassifikators verwendeten Lernverfahren problematisch ist, trachtet man danach, die Grösse des Vektorraums auf ein "reduced term set" zu beschränken. Eine solche Reduktion ist auch deshalb vorteilhaft, da sie dem "overfitting" entgegenwirkt; darunter wird die Gefahr verstanden, dass der Klassifikator eher die Spezifika der Trainingsdokumente lernt als die für die jeweiligen Klassen eigentlich konstitutiven Merkmale. Ein durch "overfitting" gekennzeichneter Klassifikator würde die Trainingsdokumente besonders gut klassifizieren können, neue Dokumente hingegen nicht. Da sich gezeigt hat, dass zur Vermeidung von "overfitting" die Zahl der Trainingsdokumente etwa proportional zur Zahl der *Terme* sein sollte, dient deren Reduktion auch der Verringerung der Zahl der benötigten Trainingsdokumente.

Die Dimensionsreduktion birgt natürlich auch die Gefahr in sich, potentiell nützliche Information zu eliminieren, weswegen verschiedene methodische Ansätze dafür entwickelt bzw. getestet wurden Bei deren Anwendung kann die Reduktion *klassenspezifisch* oder – wie zumeist – *global* erfolgen; im letzteren Fall wird eine bestimmte Un-

termenge aus allen Termen für die Klassifizierung unter allen Klassen verwendet. Die beiden im Hinblick auf die resultierenden Attribute grundsätzlich zu unterscheidenden Techniken sind die *Attributauswahl* ("feature selection", "term space reduction") und die *Attributextraktion* ("term extraction").

Attributauswahl. Dieser Ansatz trachtet danach, aus der Menge der Attribute jene herauszufiltern, die bei Verwendung als Indexterme die höchste Effizienz, d.h. die höchsten Werte hinsichtlich ihrer Bedeutung für die Klassifizierungsaufgabe erzielen.

- ***Auswahlkriterium Dokumentenhäufigkeit:*** Diese einfache und häufig verwendete Technik geht davon aus, dass nur jene Attribute verwendet werden sollen, die in der grössten Zahl von Dokumenten auftreten. Auf dieser Basis kann die Zahl der Dimensionen ohne Verlust um den Faktor 10 reduziert werden; selbst eine Reduktion um den Faktor 100 hat nur einen geringen Effizienzverlust zur Folge. Da die grosse Mehrheit der Terme in einem Dokument nur eine sehr geringe Dokumentenhäufigkeit aufweist, werden z.B. durch eine Reduktion der Attribute um den Faktor 10 jene Terme beibehalten, die über eine mittlere Dokumentenhäufigkeit (und damit den höchsten Informationsgehalt) verfügen. Varianten dieses Ansatzes sind die Entfernung aller Terme, die *nur* in *n* Trainingsdokumenten (wobei *n* oft 1, 2 oder 3 beträgt) oder die nur *n* mal im ganzen Trainingsset auftreten (wobei *n* 1 bis 5 beträgt).
- ***Informationstheoretische Auswahlkriterien:*** Eine ganze Reihe komplizierterer und informationstheoretisch begründeter Funktionen basiert auf der Annahme, dass die besten Terme für eine Klasse jene seien, die in den positiven und negativen Trainingsbeispielen am unterschiedlichsten verteilt sind.[11] Die Befunde der experimentellen Forschung deuten darauf hin, dass diese Methoden etwas bessere Resultate erzielen als die auf der Dokumentenfrequenz basierenden Selektionskriterien.

Attributextraktion. Die Attributextraktion geht davon aus, dass die in den Dokumenten auftretenden Terme aufgrund von Homonymie, Polysemie und Synonymie keine optimalen Dimensionen zur Repräsentierung der Dokumente darstellen und daher "künstliche" Terme generiert werden sollten, die von diesen Problemen nicht betroffen sind.

- ***Term Clustering:*** Mit diesem Ansatz wird versucht, Wörter mit einem hohen Grad paarweiser semantischer Verwandtschaft so zu gruppieren, dass diese Gruppen bzw. ihre Zentroide anstelle der Wörter als Dimensionen des Vektorraums verwendet werden können. Auf diese Weise sollen Synonyme bzw. inhaltlich sehr ähnliche Wörter gemeinsam repräsentiert werden. Die Cluster werden entweder durch Gruppierung von Wörtern mittels eines Ähnlichkeitsmasses, aufgrund ihres gemeinsamen

[11] Eine tabellarische Übersicht der wichtigsten (mathematischen) Funktionen findet sich bei Sebastiani (2002b, 16).

Auftretens bzw. Nicht-Auftretens in den Trainingsdokumenten oder ihrer Wahrscheinlichkeit, zur selben Klasse zu gehören, gewonnen. Bislang konnte der Clustering-Ansatz jedoch keine spektakulären Ergebnisverbesserungen bewirken.

- ***Latent Semantic Indexing:*** Diese Technik komprimiert Dokumentenvektoren in Vektoren eines geringer dimensionierten Raumes, indem die Muster des gemeinsamen Auftretens der originären Dimensionen (Terme) analysiert werden. Die neu generierten Dimensionen sind dann nicht mehr intuitiv interpretierbar, sollen aber die latente semantische Struktur der Dokumente repräsentieren. Sie müssen dabei die für eine Klasse relevantesten Begriffe nicht einmal enthalten; falls jedoch ein solcher Begriff sehr grosse Trennschärfe für die Definition einer Klasse besitzt, geht diese Information dabei verloren. Dieser Ansatz gilt als erfolgsträchtiger als der erstgenannte, insbesondere auch in Kombination mit Verfahren der *Attributauswahl.*

2.5 Induktive Erstellung von Klassifikatoren

Die induktive Erstellung eines *rangordnenden* Klassifikators für eine bestimmte Klasse besteht in der Definition einer Funktion, die bei Anwendung auf ein bestimmtes Dokument (d.h. den dieses Dokument repräsentierenden Vektor) einen Wert ausgibt, der meist zwischen "0" und "1" variiert und ausdrückt, zu welchem Grad das Dokument zu dieser Klasse gehört ("categorization status value" oder "confidence rate"). Am Ende des Prozesses werden – je nach Orientierung (vgl. Abschnitt 2.1.3) – pro Klasse die Dokumente bzw. pro Dokument die Werte für die Klassen absteigend nach diesen Werten ranggeordnet.

Die Erstellung eines *"harten"* Klassifikators basiert entweder auf einer Funktion, die eine Ja-Nein-Entscheidung trifft (das Dokument gehört zur betreffenden Klasse oder nicht), oder aber auf der Verwendung eines rangordnenden Ansatzes zuzüglich der Definition eines *Schwellenwertes* ("threshold"), der dann entscheidet, ob das Dokument zur Klasse gehört oder nicht ("categorization status value" ist grösser/gleich bzw. kleiner als der Schwellenwert).

2.5.1 Bestimmung von Schwellenwerten

Zur Bestimmung eines solchen Schwellenwerts bestehen verschiedene Möglichkeiten. Grundsätzlich ist dabei zwischen *analytischen* und *experimentellen* Ansätzen zu unterscheiden. Der analytische Ansatz ist nur bei Verwendung von Klassifikatoren möglich, die die *Wahrscheinlichkeit* der Klassenzugehörigkeit eines Dokuments vorhersagen und deren Güte durch ein entscheidungstheoretisches Mass bewertet wird ("probability thresholding"). Wenn diese theoretischen Voraussetzungen nicht gegeben sind, bedient man sich eines der folgenden experimentellen Ansätze:

- anhand einer Menge von Validierungsdokumenten werden verschiedene Schwellenwerte getestet und schliesslich jener gewählt, der die beste Klassifizierungsgüte erzielt ("categorization status value thresholding");
- man wählt jenen Schwellenwert, der die beste Übereinstimmung der Verteilung der neu zu klassifizierenden Dokumente über die Klassen mit der entsprechenden Verteilung in den Trainingsdokumenten erbringt ("proportional thresholding");
- man teilt jedem Dokument eine konstante Anzahl von Klassen zu ("fixed thresholding").

2.5.2 Arten von Klassifikatoren

Aus der Vielzahl der methodischen Ansätze sollen hier nur die wichtigsten besprochen werden:

Probabilistische Klassifikatoren. Diese berechnen auf Basis der Trainingsdokumente die bedingte Wahrscheinlichkeit der Zugehörigkeit eines Dokuments zu einer bestimmten Klasse. Da sie dabei vom Bayes'schen Theorem ausgehen und (naiverweise) annehmen, dass die Wörter in einem Dokument unabhängig voneinander auftreten, werden sie *Naive Bayes-Klassifikatoren* genannt. Diese Klassifikatoren, von denen eine Reihe von Varianten existieren, wurden und werden sehr häufig verwendet. Sie erfordern allerdings beträchtliche Mengen von Trainingsdokumenten (100 und mehr pro Klasse).

Entscheidungsbäume. Ein "decision tree classifier" ist ein regelbasiertes Verfahren, das davon ausgeht, dass jeder Knoten eine einfache Bedingung über das Auftreten eines Attributs im Text enthält; ist diese erfüllt, wird der eine Zweig des Baumes weiterverfolgt, wenn nicht, der andere. Dies wird wiederholt, bis ein "Blatt" erreicht wird, das eine Klasse darstellt, der das Dokument dann zugeordnet wird. Meist – aber nicht notwendigerweise – werden dabei nur binäre Bäume verwendet. Beim Lernen eines Entscheidungsbaumes wird geprüft, ob alle Trainingsdokumente derselben Klasse angehören; wenn dies nicht der Fall ist, wird auf der Basis eines informationstheoretischen Masses (Entropie) ein Attribut gesucht, welches die Beispieldokumente in Gruppen zerteilt, die gleiche Merkmale aufweisen; jede dieser Gruppen wird sodann zu einem separaten Sub-Baum. Der Prozess wird wiederholt, bis jedes Blatt des gesamten Baumes Trainingsdokumente enthält, die denselben Klassen angehören, für die dann diese Blätter stehen. Sebastiani (2002b, 23) unterscheidet hievon Verfahren, die *Entscheidungslisten* aufbauen ("decision rule classifiers"), die dem in Abschnitt 2.3 (Fussnote 8) erwähnten Muster folgen. Daneben gibt es eine Reihe anderer Varianten.

Diese Verfahren sind in der Erstellungs- und Trainingsphase sehr aufwendig. Ausserdem neigen sie zum "overfitting" (s.o.), d.h. sie können so spezifisch werden, dass sie nur mehr die Trainingsdokumente zu klassifizieren vermögen. Daher wird stets auch ein Mechanismus für das "Zurechtstutzen" ("pruning") der Baumstruktur benötigt. Nach Brückner (2001, 444) ist mit Entscheidungsbäumen aber auch eine hohe Güte der Klassifizierung sowie eine hohe Effizienz in der Anwendungsphase erreichbar.

Regressionsmethoden. Das bekannteste dieser Modelle ist das *Linear Least Squares Fit*-Verfahren. Bei diesem sind jedem Dokument zwei Vektoren zugeordnet, und zwar ein *Input*-Vektor aus gewichteten Termen und ein *Output*-Vektor, dessen Gewichte die Klassen repräsentieren (binär bei den Trainingsdokumenten und "categorization status values" bei den Testdokumenten). Die Aufgabe des Verfahrens ist die Bestimmung eines Output-Vektors für ein Dokument auf der Basis seines Input-Vektors, was mittels einer Matrix aus Regressionskoeffizienten für die Wörter-Klassen-Beziehungen geschieht. Dabei resultiert pro Dokument eine ranggeordnete Liste von Klassenzugehörigkeitskoeffizienten. Nach Sebastiani liegt damit einer der effizientesten Klassifikatoren vor, die heute bekannt sind; dieser Vorteil wird jedoch durch einen ausserordentlich hohen Rechenaufwand beeinträchtigt (2002b, 24).

Rocchio-Algorithmus. Dieses sehr häufig eingesetzte Verfahren[12] wurde bereits Anfang der 1970er Jahre im Rahmen des IR-Systems *SMART* (Salton & McGill 1987, 125 ff.) entwickelt. Es basiert auf der Erstellung eines prototypischen Vektors für jede Klasse auf der Grundlage von Trainingsdokumenten: Dabei wird pro Klasse den Vektoren aller zugehörigen Dokumente ein positives und den Vektoren aller nicht zur Klasse gehörigen Dokumente ein negatives Gewicht gegeben;[13] die Summe dieser Gewichte bildet dann den prototypischen Zentroid-Vektor für die Klasse. Bei der Klassifizierung eines neuen Dokuments wird die Distanz zu den Zentroid-Vektoren aller Klassen berechnet;[14] die am nächsten liegendste(n) Klasse(n) wird (werden) dann zugeordnet. Als Vorteile des Rocchio-Algorithmus gelten einfache Implementierbarkeit und geringer Rechenaufwand; Nachteile sind die stark nachlassende Güte des Verfahrens bei grossen Klassifikationssystemen (mehr als einige hundert Klassen) sowie die Schwäche des Zentroid-Ansatzes bei Klassen, deren Trainingsdokumente (zufälligerweise) untereinander Cluster irgendwelcher Art bilden.

Online-Methoden. Während es sich beim Rocchio-Algorithmus um eine lineare *Batch*-Methode, bei der alle Trainingsdokumente auf einmal analysiert werden, handelt, bilden *Online*-Methoden bald nach der Überprüfung des ersten Trainingsdokuments einen Klassifikator und modifizieren diesen schrittweise beim Überprüfen weiterer Trainingsdokumente (daher spricht man auch von *inkrementellen* Klassifikatoren). Damit eignen sie sich z.B. für Situationen, in denen nicht alle Trainingsdokumente von Anfang an verfügbar sind oder sich die Bedeutung der Klasse mit der Zeit ändert (wie etwa beim adaptiven Filtern). Sobald die echte Klasse eines Testdokuments bekannt ist (z.B. durch Benutzerfeedback), kann auch dieses zum Lernen verwendet werden. Bei diesen Algo-

[12] Bspw. bezeichnen Koster et al. den Rocchio-Algorithmus als "the typical workhorse for document retrieval and classification in IR" (2001, 20).

[13] Diese werden üblicherweise gemäss der TFIDF-Heuristik gewichtet und auf die Länge "1" normiert.

[14] Als Ähnlichkeitsmass dient meist der Kosinus des Winkels zwischen den zugehörigen Vektoren.

rithmen werden zunächst alle Gewichte des Vektors für eine Klasse auf einen Ausgangswert gesetzt; bei der Analyse eines Trainingsdokumentes wird versucht, dieses mit dem Klassenprofil zu klassifizieren. Ist das Resultat korrekt, so geschieht nichts weiter; ist es nicht korrekt, so werden die Gewichte des Klassenprofils positiv oder negativ verändert (je nach Verfahren additiv oder multiplikativ). Wenn das Dokument ein positives Beispiel war, so werden die übereinstimmenden Terme um eine Quantität ("learning rate") erhöht, wenn es ein negatives Beispiel war, werden sie verringert. Ziel ist es, eine optimale lineare Trennungsfunktion zwischen für die Klasse relevanten und nicht relevanten Dokumenten zu finden. Die am häufigsten verwendeten Modelle sind unter den Bezeichnungen *Perceptron*, *Winnow* (mit etlichen Varianten) und *Sleeping Experts* bekannt.

Künstliche neuronale Netze. Diese aus dem Bereich der Künstlichen Intelligenz entnommenen Methoden sehen ein Netzwerk aus Input-Einheiten (gewichtete Terme des Dokuments), Output-Einheiten (Klassen), sowie gewichtete Abhängigkeitsrelationen zwischen diesen vor. Beim Lernen werden diese Verbindungen auf der Basis der durch die Trainingsdokumente vorgegebenen "Lernregel" modifiziert. Das Trainieren erfolgt durch "back-propagation", d.h. das Ergebnis eines fehlgeschlagenen Versuches, ein Trainingsdokument korrekt zu klassifizieren (die echte Klasse ist ja bekannt), läuft rückwärts durch das Netz, wobei durch Modifikation der Gewichte versucht wird, die Fehler zu minimieren. Algorithmen dieser Art gelten als sehr aufwendig bei der Installation, aber auch als effizient und robust (Hoffmann 2002, 71).

Instanz-basierte Klassifikatoren. Verfahren dieser Art ("example-based classifiers") trachten nicht danach, eine optimale Repräsentation jeder Klasse aufzubauen, sondern orientieren sich an den Klassen, die jene Trainingsdokumente aufweisen, die dem zu klassifizierenden Testdokument am ähnlichsten sind.[15] Sie werden auch "lazy learners" genannt, da sie die Entscheidung, wie von den Trainingsdokumenten zu generalisieren ist, auf die Klassifizierungsphase "verschieben".

Bei der bekanntesten dieser Techniken, dem *k Nearest Neighbors-Verfahren* (*k*-NN), besteht der Lernschritt einfach darin, *alle* Trainingsdokumente zu speichern. Beim Klassifizieren eines neuen Dokuments untersucht der Algorithmus die *k* diesem Dokument ähnlichsten Trainingsbeispiele und teilt das Dokument der/den unter diesen am häufigsten vertretenen Klasse(n) zu; bei Mehrfachklassifizierung kommen Schwellenwerte zur Anwendung. Die empfohlene Anzahl für *k* liegt zwischen 20 und 45. Das Verfahren ist zwar relativ unempfindlich gegenüber "verrauschten Trainingsbeispielen", da diese von die Mehrheit ihrer nicht-verrauschten Nachbarn dominiert werden (Klinken-

[15] Dafür werden verschiedene Ähnlichkeits- oder Distanzmasse, am häufigsten jedoch der Kosinus zwischen den Dokumentenvektoren, herangezogen.

berg 1998, 30), weist aber einige Verzerrungsgefahren auf (Hoffmann 2002, 66f.) Die Geschwindigkeit des Trainings wird durch einen beträchtlichen Aufwand in der Klassifizierungsphase erkauft, da in dieser die gesamte Trainingskollektion nach der Ähnlichkeit jedes Dokuments mit dem Testdokument ranggeordnet werden muss (Sebastiani 2002b, 29).

Support-Vektor-Maschinen. Dieser erst Ende der 1990er Jahre eingeführte Ansatz gilt als besonders vielversprechend, da damit auffallend gute Testergebnisse erzielt wurden. "Support vector machines" (SVM) ist ein komplexes mathematisches Verfahren, das der Mustererkennung entstammt und sich für binäre Klassifizierungsprobleme einsetzen lässt. Dabei wird eine "Hyperebene" berechnet, die die positiven und negativen Trainingsbeispiele optimal trennt. Aus dem Training resultieren für jede Klasse die meist nur wenigen Trainingsvektoren, die am nächsten zu dieser Hyperebene liegen; sie werden "Support-Vektoren" genannt und gelten als die einzigen wirkungsvollen Elemente der Trainingsmenge – würden alle anderen Trainingsdokumente entfernt, so würde der Algorithmus immer noch dieselbe Entscheidungsfunktion lernen (Yang & Liu 1999, 44). Die Klassifizierung wird im wesentlichen durch die Ermittlung der Support-Vektoren mit der kürzesten Distanz zum jeweiligen Testdokument vorgenommen (Brückner 2001, 445). Der Vorteil von SVM ist, dass dieses Verfahren oft keine Dimensionsreduktion erfordert, da es sehr grosse Attributzahlen bewältigen kann und dabei robust gegenüber dem "overfitting" ist; in jüngster Vergangenheit wurden auch sehr leistungsfähige Algorithmen für SVM entwickelt, sodass der früher festgestellte hohe Rechenaufwand kein Hinderungsgrund für die Verwendung des Verfahrens mehr ist (Sebastiani 2004, Kpt. 3.2.1).

2.5.3 Kombination von Klassifikatoren

Hinter der Idee eines kombinierten Einsatzes von Klassifikatoren ("classifier committees", "classifier ensembles"), steht die These, dass "*k* different classifiers [...] may be better than one if their individual judgements are appropriately combined" (ibid., Kpt. 3.2.2). In diesem Zusammenhang geht es einerseits um die Auswahl der betreffenden Verfahren und andererseits um die Definition einer geeigneten Kombinationsfunktion. Bezüglich des ersten Kriteriums gilt, dass die kombinierten Verfahren so unabhängig bzw. unterschiedlich wie möglich sein sollten. Im Hinblick auf den zweiten Aspekt wurden mehrere Regeln getestet (Mehrheitsentscheidung, gewichtete lineare Kombination usw.) Die bisher vorliegenden Resultate lassen jedoch keine eindeutigen Schlüsse zu (Sebastiani 2002b, 31).

Als Spezialfall des Kombinierens von Klassifikatoren gilt "boosting", ein Ansatz, bei dem es sich bei den "kombinierten" Verfahren stets um denselben Klassifikator handelt. Dabei erfolgt eine *sequentielle* Vorgangsweise, d.h., dass das Verfahren beim

wiederholten Einsatz die Resultate der früheren Versuche berücksichtigen und insbesondere versuchen kann, jene Fälle zu verbessern, die seine "Vorgänger" am schlechtesten gelöst haben. Dieser Ansatz ist relativ neu und gilt als sehr erfolgversprechend (Sebastiani 2004, Kpt. 3.2).

2.6 Evaluierung der Klassifizierungsgüte

Bei der experimentellen Evaluierung eines automatischen Klassifizierungsverfahrens wird üblicherweise dessen Wirksamkeit ("effectiveness") gemessen, d.h. die Fähigkeit, die *richtigen* Klassifizierungsentscheidungen zu treffen. Daneben kann auch die Leistungsfähigkeit ("efficiency") eines Verfahrens bewertet werden, d.h. die durchschnittliche Geschwindigkeit bei der Konstruktion eines Klassifikators auf Basis der Trainingsdokumente bzw. bei der Klassifizierung eines neuen Testdokuments. Hier soll nur der erste Aspekt – die Bewertung der *Klassifizierungsgüte* – näher betrachtet werden.

2.6.1 Masse für die Klassifizierungsgüte

Precision und Recall. Diese beiden aus dem IR wohlbekannten Kriterien werden auch bei der Bewertung von Ergebnissen des automatischen Klassifizierens herangezogen. In diesem Zusammenhang werden diese Masse wie folgt definiert:

- *Precision* wird als die Wahrscheinlichkeit verstanden, mit der bei Zuordnung eines (zufälligen) Dokuments zu einer bestimmten Klasse diese Entscheidung korrekt ist (Grad der Zuverlässigkeit);
- *Recall* bezeichnet die Wahrscheinlichkeit, dass, wenn ein (zufälliges) Dokument einer bestimmten Klasse zugeordnet werden müsste, dies auch geschieht (Grad der Komplettheit).

Zuordnung zu Klasse N		**Richtige Klasse (Expertenurteil)**	
		Ja	Nein
Automatisches Verfahren	Ja	A	B
	Nein	C	D

Tabelle 2-1: Kontingenztabelle für ein Klassifizierungsproblem (nach Klinkenberg 1998, 17; Sebastiani 2002b, 33)

Dies kann für jede Klasse mit Hilfe einer *Kontingenztabelle* dargestellt werden (*Tabelle 2-1*), in der die Resultate des automatischen Verfahrens der korrekten Klassifizierung, die zuvor durch Experten vorgenommen wurde, gegenübergestellt werden. Dabei bezeichnen die Zelleninhalte folgende Ergebnisse:

A = vom Verfahren korrekt klassifiziert ("true positives")

B = vom Verfahren fälschlich zugeordnet ("false positives"; "errors of commission")

C = vom Verfahren fälschlich nicht zugeordnet ("false negatives"; "errors of omission")

D = vom Verfahren korrekterweise nicht zugeordnet ("true negatives")

Anhand der Tabelle lassen sich die beiden Masse wie folgt ermitteln:

$$\text{Precision} = A / (A + B)$$

$$\text{Recall} = A / (A + C)$$

Für die Berechnung der Durchschnittswerte für Precision und Recall beim automatischen Klassifizieren eines Dokumentenkorpus (nach mehr als zwei Klassen) gibt es zwei Methoden:

- beim "microaveraging" wird eine *globale* Kontingenztabelle erzeugt, indem die klassenspezifischen Kontingenztabellen aufsummiert werden;
- beim "macroaveraging" werden die beiden Parameter zunächst klassenspezifisch berechnet und erst dann gemittelt.

Nach Sebastiani (2002b, 33; 2004, Kpt. 2.3) können diese beiden Berechnungsmethoden sehr unterschiedliche Ergebnisse liefern. Die erste "belohnt" Verfahren, die bei stark besetzten Klassen gut funktionieren, die zweite hingegen solche Verfahren, die auch bei schwach besetzten Klassen gut funktionieren.

Accuracy.[16] Ein weiteres Gütekriterium ist das Mass *Accuracy*, das die Wahrscheinlichkeit ausdrückt, dass ein zufällig ausgewähltes Dokument korrekt klassifiziert wird. Diese lässt sich abschätzen, indem die Anzahl der korrekten Zuordnungen durch die Anzahl aller Zuordnungen dividiert wird:

$$\text{Accuracy} = (A + D) / (A + B + C + D)$$

Manchmal wird auch die als *Error* (Fehlerrate) bezeichnete Gegenwahrscheinlichkeit betrachtet:

$$\text{Error} = 1 - \text{Accuracy}$$

Nach Klinkenberg (1998, 17) ist Accuracy (bzw. Error) das im maschinellen Lernen am häufigsten verwendete Performanzmass für Lernverfahren. Sebastiani (2002b, 34) betont jedoch, dass beim automatischen Klassifizieren die Verwendung von Accuracy weniger günstig sei, da das Mass durch den hohen Wert im Nenner weniger sensitiv auf Variationen der Zahl der korrekten Zuordnungen (A+D) reagiert als Precision und Recall.

Utility. Dieses Mass gilt als Alternative zu Gütekriterien, da es diese um ökonomische Kriterien wie Gewinn oder Verlust erweitert. Analog zur Kontingenztabelle wird eine Nutzentabelle erstellt, in der numerische Werte den Gewinn durch den jeweiligen Fall

[16] In manchen Untersuchungen wird zwischen Accuracy und Precision nicht ausreichend unterschieden; bspw. setzen Prabowo et al. (2002b, 1) die beiden Masse gleich, meinen jedoch Accuracy.

(A, B, C, D wie in *Tabelle 2-1*) ausdrückt. Als Anwendungsbeispiel führt Sebastiani (2004, Kpt. 2.3) "spam filtering" an, eine binäre Klassifizierungsaufgabe, bei der Precision wichtiger ist als Recall, da die Einstufung einer legitimen E-Mail-Nachricht als "spam" einen schlimmeren Fehler darstellt als die Einstufung einer Spam-Nachricht als "non-spam".

Gütekriterien bei der hierarchischen Klassifizierung. Frank & Paynter (2004, 222f.) erwähnen vier Kriterien, die im Falle einer hierarchischen Klassifizierung aussagekräftiger sein sollen als das übliche Accuracy-Mass:

- *Percent too specific* misst den Anteil der zugeteilten Klassen, die im Vergleich zu den "wahren" Klassen zu spezifisch sind (der korrekten Pfad durch die Hierarchie ist ein Präfix des durch den Klassifikator vorhergesagten);
- *Percent too general* misst den Anteil der zugeteilten Klassen, die im Vergleich zu den "wahren" Klassen zu wenig spezifisch sind (der vorhergesagte Pfad ist ein Präfix des korrekten Pfades);
- *Average overlap* misst das Ausmass der Überschneidungen zwischen dem vorgesagten und dem korrekten Pfad; dies wird berechnet, indem die Zahl der übereinstimmenden Verzweigungen ("nodes") zur Maximallänge des Pfades in Beziehung gesetzt wird;
- *Accuracy at level* drückt auf jeder Hierarchiestufe den Anteil der korrekten Zuteilungen aus.

Kombinierte Gütekriterien. Daneben gibt es verschiedene Ansätze, um die Bewertung (wie im Fall von Accuracy) auf eine einzelne Zahl zu bringen:

- ***Eleven-point average precision:*** Durch Veränderung des Schwellenwerts wird der Recall auf die Werte "0,0", "0,1", ..., "0,9", "1.0" gesetzt, für diese 11 Fälle die Precision ermittelt und dann der Durchschnitt gebildet.
- ***Precision-recall breakeven point:*** Dies ist der Punkt, bei dem Precision und Recall gleich sind; er wird durch einen ähnlichen Prozess wie oben ermittelt, wobei meist ein Plot der Precision als Funktion des Recalls erstellt und der Schnittpunkt gesucht bzw. interpoliert wird.
- ***F-Mass** oder **F-Funktion:*** Dabei werden Precision (P) und Recall (R) in einem einzigen Mass kombiniert, wobei meist von gleicher Bedeutung (="1") der beiden Parameter ausgegangen wird ("F1-Mass"):[17]

$$F1 = \quad 2 * P * R / (P + R)$$

Nach der Auswahl eines Gütekriteriums wird meist ein "tuning" des Klassifikators vorgenommen, um durch Variierung eines Parameters (z.B. eines Schwellenwertes) das bestmögliche Resultat zu erzielen. Dies muss experimentell geschehen, d.h. durch wiederholtes Klassifizieren der Validierungsdokumente mit veränderten Werten nur die-

[17] Auf die Darstellung der komplizierteren allgemeinen Formel, die ein Gewicht für die *Bedeutung* von Precision / Recall enthält, kann daher hier verzichtet werden.

ses Parameters. Der Wert, der dabei die beste Klassifizierungsgüte erreicht, wird sodann bei der Klassifizierung neuer Dokumente verwendet.

2.6.2 Benchmarks

Für das Experimentieren mit Klassifikatoren usw. gibt es eine Reihe von öffentlich zugänglichen Benchmark-Korpora bereits klassifizierter Dokumente, die zahlreichen Untersuchungen verwendet wurden. Die bekanntesten sind:

- Testkollektion *Reuters-21578*[18]: Mit diesem Textkorpus (bestehend aus Artikeln von Nachrichtenagenturen) bzw. seinen Vorgänger-Versionen (*Reuters-22173* u.a.) wurde der grösste Teil aller Experimente zur automatischen Textklassifizierung durchgeführt (Sebastiani 2002b, 37). Im Hinblick auf die Klassifizierung von Web-Ressourcen erscheint dieses Korpus jedoch als eher problematisches Modell, da es vergleichsweise "small and tidy" ist und überdies die "Klassifikation" nur 135 Kategorien (mit Wirtschafts- und Politikbezug) ohne hierarchische Struktur umfasst (Dumais & Chen 2000, 258). Einige Forscher haben daher selbst (bescheidene) Hierarchien hinzugefügt (ibid.)
- Testkollektion *OSHUMED:* Enthält Titel und z.T. Abstracts aus medizinischen Zeitschriften, die, da sie der Datenbank *MEDLINE* entnommen wurden, mit *MeSH*-Deskriptoren versehen sind.
- Testkollektion *20 Newsgroups*: Enthält nach diesen 20 Newsgruppen kategorisierte Usenet-Artikel.

Sebastiani (2002b, 37) hebt hervor, dass bei der Verwendung dieser oder anderer Standardkollektionen, die dem methodischen Vergleich dient (z.B. für "cross-classifier" Vergleiche), korrekterweise drei Bedingungen erfüllt sein müssen. Dies sind die Verwendung genau derselben Kollektion (Dokumente und Klassen), derselben Aufteilung zwischen Trainings- und Testmenge sowie desselben Gütekriteriums. "Unfortunately, a lot of experimentation, both on Reuters and on other collections, has not been performed with these three *caveats* in mind" (ibid.)

2.6.3 Suche nach dem besten Klassifikator

Ein grosser Teil der (Informatik-)Literatur zum Thema "automatisches Klassifizieren" widmet sich dennoch der Frage, welcher Klassifikator in Kombination mit welchen Randbedingungen (d.h. vielen der in diesem Kapitel besprochenen Aspekte) die beste Leistung erbringe. Mit der nötigen Vorsicht zog Sebastiani vor wenigen Jahren die folgenden Schlussfolgerungen (2002b, 39–40):

- die besten Resultate wurden durch Kombination von Klassifikatoren mittels "boosting", durch Support-Vektor-Maschinen, Instanz-basierte Klassifikatoren und Regressionsmethoden erzielt, ohne dass aber eines dieser Modelle als "Spitzenreiter" bezeichnet werden kann;
- neuronale Netze und Online-Klassifikatoren stehen eine Stufe darunter;
- Rocchio- und Naive Bayes-Verfahren scheinen schlechtere Leistungen zu erbringen;

[18] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html bzw. http://www.daviddlewis.com/resources/testcollections/reuters21578/ [beide: 28.06.2004]

- die mit Abstand schlechteste Leistung wurde für den Klassifikator *WORD* gemessen, ein nur zu Vergleichszwecken implementiertes, nicht-lernendes Verfahren, das lediglich auf einem Vergleich zwischen den Dokumenten und den Klassenbenennungen beruht, die als gewichtete Terme im Vektorraummodell betrachtet werden.[19]

Zunehmend populärer scheint die Idee zu werden, die hypertextuelle Struktur des WWW für eine Verbesserung der Klassifizierung mitzunutzen. Dies bedeutet, dass entweder die Inhalte der in einem vorliegenden Web-Dokument durch Hyperlinks *zitierten* Dokumente oder die Inhalte der das gegebene Web-Dokument mittels Hyperlinks *zitierenden* Dokumente (oder auch die Kombination von beidem) *zusätzlich* zum Vokabular des Vorlagedokuments oder sogar *anstelle* dessen im Klassifizierungsprozess verwendet werden. Mehrere Studien berichten von z.T. deutlich besseren Klassifizierungsergebnissen unter Verwendung dieses Ansatzes (z.B. Attardi et al. 1999; Fürnkranz 1999; Glover et al. 2002; Kuo & Wong 2000; Oh et al. 2000).

2.7 Labor-, Open Source und kommerzielle Software

Angesichts der Erfolge der im Experiment verwendeten Verfahren – der grösste Teil der Literatur zum automatischen Klassifizieren bezieht sich ja auf diese – ist es nicht überraschend, dass inzwischen eine ganze Fülle von kommerziellen Softwareprodukten existiert, die auf den im Labor getesteten Verfahren basieren und für sich in Anspruch nehmen, Klassifizierungsprobleme vielfältiger Art – vor allem jedoch die im Rahmen des betrieblichen Informationsmanagements auftretenden – lösen zu können. Im Hinblick auf komplexe Klassifikationssysteme und grosse Datenmengen mag diesbezüglich jedoch eine gewisse Skepsis angebracht sein (Krier & Zaccà 2002, 190). "These automated classification tools work with varying degrees of success" (Hagedorn 2001, 5).

Diese Produkte, auf die im Rahmen dieser Arbeit nicht im Detail eingegangen werden kann, unterscheiden sich beträchtlich hinsichtlich der eingesetzten Methodik sowie anderer Kriterien. Man kann die Produkte z.B. danach unterscheiden, ob sie eine Clusterlösung, eine Klassifizierung auf der Basis eines vorhandenen (hierarchischen) Schemas – in diesem Kontext zumeist als "Taxonomie" bezeichnet – oder beides anbieten.

Walther (2001) versuchte einen Vergleich von neun Anbietern mittels Fragebogen bzw. Firmenliteratur. Hoffmann (2002, 85–90) bietet aussagefähigere Details zu acht Produkten. Ein aktueller, wenngleich nur kurzer tabellarischer Vergleich von sieben Anbietern findet sich bei Adams (2003).

[19] In den nächsten Kapiteln wird sich zeigen, dass die grossen und bekannten Projekte zum automatischen Klassifizieren (mit bibliothekarischen Schemata) z.T. auf methodischen Ansätzen beruhten, die diesem stark ähneln.

Eine umfangreiche Marktstudie zu diesem Bereich legte vor zwei Jahren das Beratungsunternehmen Delphi Group (Boston, MA) vor (Delphi Group 2002). In diesem Bericht wird auf die Produkterwartungen von 450 befragten Unternehmungen, die Struktur des Marktes sowie auf elf Anbieter/Produkte im Detail eingegangen. Eine willkürliche Auswahl[20] einiger Anbieter und Produkte zeigt *Tabelle 2-2*:

Anbieter	Produkt	Literatur
AmikaNow!	AmikaClassifier	AmikaNow! 2002a; 2002b
Autonomy	Autonomy Classification	Autonomy o.J.
Brainbot	nextbot Klassifikator	Hees 2004
Convera	RetrievalWare	Convera 2002
Insuma	Insuma Distributed Search Engine	Insuma 2002
Intology	Klarity; Taxonomy Builder	Intology o.J. a; o.J. b
Inxight	Inxight SmartDiscovery	Inxight 2004
Kofax	Moho Classifier	Kofax 2004
Recommind	MindServer	Recommind 2004
7d	7d classify	Sieben-D 2003
Thunderstone	Texis Categorizer	Thunderstone o.J.
Verity	Verity Intelligent Classifier	Verity 2001

Tabelle 2-2: Einige Anbieter kommerzieller Klassifizierungssoftware

Über erste Erfahrungen mit den beiden kommerziellen Produkten *Intelligent Miner for Text* (IBM) und *Brainware* (SER) berichtete Renz (2001); danach wurden – allerdings nur im didaktischen Umfeld und mit kleinen Datenmengen – relativ zufriedenstellende Resultate erzielt. Die Anwendung verschiedener kommerzieller Produkte in Unternehmungen beschreiben z.B. Blumberg & Atre (2003) und Lamont (2003). Auch im Bereich der Mediendokumentation beginnen solche Produkte Fuss zu fassen; vgl. z.B. den Test von Kleinoeder & Puzicha (2002) beim ZDF oder die Lösung beim Süddeutschen Verlag (vgl. Abschnitt 7.3). Schliesslich sei auf das gegenwärtig in Entwicklung befindliche Projekt *GERHARD II* hingewiesen, im Rahmen dessen ebenfalls eine kommerzielle Klassifizierungssoftware eingesetzt werden soll (vgl. Abschnitt 5.3).

Die Preise für kommerzielle Klassifizierungssoftware sind relativ hoch. So muss z.B. für den Erwerb des *Texis Categorizer* mit rund $10.000,- und nochmals derselben

[20] Dabei handelt es sich um jene Anbieter/Produkte, für die im Rahmen der Recherchen für die vorliegende Arbeit Firmenliteratur angefallen ist.

Summe für die benötigte Zusatzsoftware gerechnet werden, was z.T. noch weit unter den Preisen mancher Mitbewerber liegt (Rapoza 2002).[21]

Verfügt man über keinen Zugang zu kommerzieller Software, möchte aber dennoch mit Verfahren zur automatischen Klassifizierung arbeiten, bietet sich *Open Source* Software an. Auch diese existiert in überraschend grosser Zahl. So bietet etwa OCLC seit 2002 eine Open Source-Version des hauseigenen Klassifizierungsverfahrens *Scorpion* an (vgl. Abschnitt 6.9.3). Hoffmann (2002, 91 ff.) liefert Details zu den fünf Open Source-Paketen *libBow,*[22] *LNKnet,*[23] *NL ToolKit,*[24] *SVM*^{light} [25] und *WEKA,*[26] die jeweils einen oder mehrere verschiedene Klassifikationsalgorithmen (Naive Bayes, *k*-NN, Rocchio, SVM usw.) für verschiedene Plattformen von Betriebssystemen enthalten. Eine grössere Zahl von Programmen lässt sich über die Web-basierten (Open-Source-)Software-Register *freshmeat*[27] und *SourceForge*[28] finden, wobei in beiden Fällen die Suchbegriffe "classifier" bzw. "classification" zum Erfolg führten. Weitere reichhaltige Quellen sind die Seite "Software for Classification" des Web-Dienstes *KDnuggets,*[29] der Verzeichniszweig "Top: Computers: Software: Databases: Data Mining: Public Domain Software" des Web-Katalogs *DMOZ*[30] sowie die umfangreiche japanische Linksammlung "Software Tools for NLP".[31]

[21] Als Preise für zwei weitere, allerdings auch etwas umfangreichere Produkte gibt Rapoza (2002) Grössenordnungen von $85.000,-/110.000,- bzw. $140.000,-/160.000,- an.

[22] http://packages.debian.org/stable/text/libbow [06.07.2004]

[23] http://www.ll.mit.edu/IST/lknet/index.html [06.07.2004]

[24] http://nltk.sourceforge.net/ [06.07.2004]

[25] http://svmlight.joachims.org/ [06.07.2004]

[26] http://www.cs.waikato.ac.nz/~ml/weka/ [06.07.2004]

[27] http://freshmeat.net/ [07.07.2004]

[28] http://sourceforge.net/ [07.07.2004]

[29] http://www.kdnuggets.com/software/classification.html [07.07.2004]

[30] http://dmoz.org/Computers/Software/Databases/Data_Mining/Public_Domain_Software/ [07.07.2004]

[31] http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp_tools.html [07.07.2004]

3 Die Projekte an der Universität Lund

NetLab ist eine Anfang der 1990er Jahre an der Universitätsbibliothek Lund (Schweden) entstandene multidisziplinäre Forschungs- und Entwicklungsabteilung, die sich vor allem mit den Themenbereichen "digitale Bibliothek" und "globale vernetzte Informationsdienste" beschäftigt (Ardö et al. 2002). Im Laufe des vergangenen Jahrzehnts realisierte NetLab zwei Projekte, im Rahmen derer Verfahren des automatischen Klassifizierens erprobt und zum Teil auch in Web-basierten Lösungen eingesetzt wurden. NetLab kooperierte dabei auch mit den weiter unten besprochenen Projekten der Universität Oldenburg und von OCLC. Im folgenden werden die beiden NetLab-Projekte sowie ein drittes, aktuelleres Projekt der UB Lund analysiert.

3.1 Nordic WAIS / WWW[1]

Ziel des ersten Projekts, *Nordic WAIS / World Wide Web* (auch *W4*; durchgeführt von Sommer 1993 bis Sommer 1994 gemeinsam mit der Nationalen Technischen Bibliothek Dänemarks), war die Verbesserung der damals existierenden Internet-Suchdienste und insbesondere die Integration von WAIS[2] und WWW auf der Basis wechselseitiger Gateway-Lösungen. Hinsichtlich WAIS wurde als eines der wichtigsten Probleme für die Benutzung die völlig unzureichende Präsentation der etwa 700 Datenbanken hinsichtlich ihres Inhalts empfunden. Dieser Mangel sollte durch die Erstellung einer Browsingfähigen Baumstruktur ("subject tree") im WWW behoben werden, die diese Datenbanken auf der Basis eines etablierten bibliothekarischen Klassifikationssystems erschliessen würde.

3.1.1 Methodische Vorgangsweise

Für jede WAIS-Datenbank wurde eine Wortliste erstellt, die sich einerseits auf die Datenbankbeschreibungen aus dem "directory of servers" ("keywords", d.h. freie Schlagwörter, sowie Freitext) und andererseits auch auf andere Beschreibungen (fachliche Go-

[1] Literatur zu diesem Projekt (mit einem beträchtlichen Grad von Wiederholungen/Überschneidungen): Ardö et al. (1994a; 1994b); Ardö & Koch (1993; 1994); Koch (1994); Koch et al. (1995; 1997, 32–33).

[2] WAIS (Wide Area Information Server) war eine frühe Implementierung des Z39.50-Standards am Internet. Der Dienst ermöglichte die Recherche in fachlich spezialisierten Datenbanken, die auf verteilten Servern gehalten und in einem "directory of servers" nachgewiesen wurden, mittels eines WAIS-Klienten und sah u.a. eine einfache natürlichsprachige Abfrage sowie Relevanz-Ranking und -Feedback vor. Das System wurde jedoch mit dem Siegeszug des WWW obsolet und ist heute praktisch nicht mehr existent. (http://en.wikipedia.org/wiki/Wide_area_information_server [29.04.2004])

pher-Dienste[3] und Datenbanklisten) stützte. Die Wörter wurden nach ihrer Herkunft in vier Gruppen eingeteilt (Ardö & Koch 1994):

- words from the description field
- words from the description field marked as keywords together with the name of the database
- words from the keyword-list field
- words from the subjects field

Aufgrund der resultierenden Wortlisten wurde pro Datenbank eine Liste potentiell zutreffender Notationen erstellt, indem die Wörter mit dem Vokabular der aus der UDC (English Medium Edition) verglichen wurden.[4] Das UDC-Vokabular entstammte allerdings nur einem Teil dieses Klassifikationssystems und auch nur den beiden obersten Hierarchiestufen (insgesamt 51 Klassen). Bei Übereinstimmung ("match") wurde die entsprechende Notation auf die Liste der Kandidaten-Notationen gesetzt, wobei eine Gewichtung nach den oben genannten Gruppen erfolgte (z.B. erhielten Wörter aus dem Feld "subject" ein höheres Gewicht als solche aus dem Feld "description"). Am Ende des Klassifizierungsprozesses wurden mithilfe eines heuristischen Verfahrens, das die summierten Gewichte pro Notation sowie die Zahl der vorgeschlagenen Notationen in Betracht zog, die endgültigen Notationen festgelegt, die sodann für die Einordnung der betreffenden Datenbank in die WWW-Baumstruktur massgeblich waren.

3.1.2 Evaluierung

Die Ergebnisse des Verfahrens wurden offenbar nur durch intellektuelle Inspektion evaluiert. Die Zahl der Fehlklassifizierungen wurde mit "erstaunlich gering" angegeben. Etwa 10% der Datenbanken erhielten wegen mangelnder Übereinstimmung des Vokabulars überhaupt keine Zuordnung; dies beruhte auf dem Fehlen signifikanter Stichwörter in der Beschreibung sowie dem limitierten UDC-Vokabular. Die von NetLab ins Auge gefassten Erweiterungen des Vokabulars (Beschreibungen und UDC) wurden jedoch offensichtlich nie realisiert.

3.1.3 Benutzung

Der WAIS-Index wurde als Teil der "Lund University Electronic Library" angeboten und von Benutzern aus 50 Ländern verwendet, wobei wöchentliche Transaktionszahlen von 6.000 (Ardö et al. 1994b) bzw. 15.000 (Koch 1994) genannt wurden. Seit Anfang 1996 wurde der Dienst jedoch im Zuge des Niedergangs von WAIS eingefroren, da kein "directory of servers" mehr zur Verfügung stand und es nicht mehr möglich schien, In-

[3] Gopher ist ein weiteres damals übliches und inzwischen obsolet gewordenes Internet-Protokoll.

[4] Warum gerade die UDC herangezogen wurde, wurde in der Literatur nicht begründet – vielmehr findet sich bloss (mehrfach) der Hinweis, dass man genauso gut die LCC hätte verwenden können.

formationen über die inzwischen hinter WWW-Seiten verborgenen WAIS-Datenbanken zu erhalten (Koch et al. 1997, 33). Die Homepage des Dienstes mit einer alphabetischen Auflistung der Datenbanken sowie der UDC-basierten Baumstruktur kann aber immer noch aufgerufen werden.[5]

3.2 DESIRE II

DESIRE war ein in zwei Phasen von NetLab gemeinsam mit einer Reihe weiterer Partner[6] durchgeführtes EU-Projekt. Die erste Phase konzentrierte sich auf die Analyse und Konzeption einer Informations-Infrastruktur für die europäische akademische Benutzergemeinschaft[7] (Koch 1997b) und beinhaltete im Hinblick auf das automatische Klassifizieren lediglich eine Analyse von etablierten Klassifikationssystemen und deren Eignung für die Erschliessung und Recherche von Internet-Ressourcen (Koch et al. 1997), einschliesslich eines kurzen Überblicks über bisherige Versuche zur automatischen Klassifizierung solcher Dokumente (ibid., 32–35; Koch & Day 1998). Die zweite Phase, *DESIRE II* (ab Juli 1998), die auf die Verbesserung und Weiterentwicklung von SBIGs sowie regionaler, automatisch erstellter, Metadaten-basierter Web-Archive abzielte, sah hingegen "automatic classification" als einen der Arbeitsschwerpunkte vor ("task 3.6a"), der drei Schritte umfassen sollte (Koch 1999; Zettergren 2000, 3):

- State-of-the-art report on projects, methods, alternatives and problems with automatic classification.[8]
- Tests using several automatic classification methods on the "All" Engineering (AE) robot-generated database of engineering documents from the Internet.
- Pilot service, demonstrator: "All" Engineering with classification/browsing option.[9]

[5] http://www.lub.lu.se/W4/ bzw. http://www.lub.lu.se/auto_new/UDC.html [02.03.2004]

[6] Von diesen sind ILRT (Institute for Learning and Research Technology, University of Bristol, UK) und UKOLN (UK Office for Library and Information Networking, University of Bath, UK) hervorzugeben.

[7] So z.B. durch die Analyse bestehender Internet-Suchdienste (Koch et al. 1996), die Erstellung eines prototypischen "Nordic Web Index" (Ardö & Lundberg 1998) sowie die Entwicklung der Harvesting-Software COMBINE (http://www.lub.lu.se/combine [15.05.2004]), die in der zweiten Phase weiter verbessert wurde (Cross 2000).

[8] Ob dieser Bericht jemals veröffentlicht wurde, konnte im Zuge der Recherchen zur vorliegenden Arbeit nicht geklärt werden.

[9] Die Literatur zu diesem Arbeitsschwerpunkt ist schwer überblickbar, da sie sich nicht nur durch eine grosse Zahl von Publikationen und ein beträchtliches Ausmass an Duplizierung auszeichnet, sondern – zumindest was die Web-Dokumente darunter betrifft – einen hohen Grad von wechselseitigen Verlinkungen aufweist, die allerdings z.T. nicht mehr funktionsfähig sind; letzteres betrifft insbesondere die Demonstrationsseite (Ardö & Koch 2000) und eine umfangreiche Linksammlung (Koch 2000). Soweit nicht anders angegeben, stützt sich die folgende Darstellung auf zwei Zusammenfassungen (Koch & Ardö 2000a; Koch et al. 2000), auf einen auf der Londoner Online-Tagung 1999 gehaltenen Vortrag (Ardö & Koch 1999a) sowie vor allem auf die Arbeitspapiere 1 und 2 des Projekts (Ardö et al. 1999; Koch & Ardö 2000b).

3.2.1 Engineering Electronic Library, Sweden, und "All" Engineering[10]

Die Tests wurden am Beispiel des Fachgebietes "Ingenieurwissenschaften" vorgenommen. Bereits 1994 war von NetLab im Auftrag der schwedischen technischen Universitäten der fachbezogene Qualitätsdienst *EELS* (Engineering Electronic Library, Sweden) eingerichtet worden, der etwa 1.400 ausgewählte und qualitativ hochwertige ingenieurwissenschaftliche Internet-Ressourcen beinhaltete, die von Fachreferenten und Experten intellektuell mittels des Thesaurus und des Klassifikationssystems der Datenbank *COMPENDEX* von *Engineering Information, Inc. (Ei)*[11] erschlossen und auf der Basis dieses Klassifikationssystems für das Browsing am Web aufbereitet wurden.[12]

Zu diesem SBIG trat ab 1996 die Datenbank *AE* ("All" Engineering), die die relativ kleine Zahl der in *EELS* nachgewiesenen Quellen quantitativ erweitern sollte und rund 127.000 (später 253.000) mit dem Harvester COMBINE gesammelte Web-Dokumente umfasste. Diese waren zwar über Volltext, Titel, Headings usw. suchbar, verfügten aber über keine sachliche Erschliessung.[13] Bereits im Projekt *DESIRE I* war auf Basis von Z39.50 und Dublin Core ein Interface eingerichtet worden, welches die simultane Recherche in *EELS* und *AE* erlaubte.[14] Unter *DESIRE II* wurden verschiedene Harvesting-Strategien für *AE* getestet und verglichen (Ardö et al. 1999, 3–7), was im Hinblick auf das eingesetzte automatische Klassifizierungsverfahren insoferne von Bedeutung war, als dieses eine fachlich einigermassen homogene Ausgangsmenge von Dokumenten erforderte. Die Idee, schon an dieser Stelle einen Informationsfilter – also eine in die Harvesting-Software integrierte automatische Klassifizierungskomponente zur Vorauswahl relevanter Dokumente – vorzusehen, wurde bereits zu diesem Zeitpunkt geäussert, jedoch nicht realisiert. Wie die Überprüfung von Stichproben zeigte, erbrachten die beiden in die engere Wahl gezogenen Harvesting-Ansätze jeweils rund 200.000 Datensätze mit etwa 77% fachlich relevanten Dokumenten. Da die sich die beiden Datenmengen aber nur geringfügig überschnitten, lag es nahe, die Methoden zu kombinieren (Ardö & Koch 1999b). Als Testdatenbank wurden schliesslich 155.611 Dokumente von 28.500 verschiedenen Hosts mit durchschnittlich 10 Links pro Dokument herangezogen.

Die Herausforderung für *DESIRE II* bestand nun darin, den Qualitätsdienst *EELS* mit der roboter-generierten und nicht intellektuell erschlossenen Datenbank *AE* auf Basis einer gemeinsamen Browsing-Struktur zu kombinieren ("cross-browsing"). Für die letztere sollte eine solche Struktur mittels eines automatischen Klassifizierungsverfahrens erstellt werden.

---

[10] Vgl. Ardö et al. (1999)

[11] http://www.ei.org/ [20.06.2004]

[12] Vgl. http://eels.lub.lu.se/ [26.02.2004]. Dieser Dienst wurde 2001 eingestellt, ist aber auf dem damaligen Stand weiterhin zugänglich.

[13] Vgl. http://eels.lub.lu.se/ae/index.html [01.03.2004]. Auch dieser Dienst wurde inzwischen eingestellt.

[14] Vgl. http://eelsdb.lub.lu.se/aeels/search.html [15.05.2004]

3.2.2 *Ei*-Klassifikation und *Ei*-Thesaurus

Die für das automatische Verfahren verwendete *Ei-Klassifikation* weist etwa 800 Klassen mit fünf Hierarchiestufen auf (sechs Hauptklassen, 38 Unterklassen, 182 Klassen auf der dritten Ebene usw.) Als Vokabular für das Klassifizierungsverfahren stand zunächst der Text der Klassenbenennungen zur Verfügung. Um dieses limitierte Vokabular zu erweitern, wurde neben dem Klassifikationssystem auch der *Ei-Thesaurus*[15] herangezogen, der insgesamt rund 17.500 Einträge (davon fast 8.300 "preferred terms", d.h. Deskriptoren) verzeichnet, die vom Hersteller intellektuell klassifiziert (als "main class" oder als "optional class") worden waren. Bei den Thesaurus-Einträgen handelte es sich um Einzelwörter und Mehrwortgruppen ("composite terms") mit vorwiegend 2 bis 3 Wörtern.

Im Zuge der Vor- oder Aufbereitung ("preprocessing") wurde eine Konkordanz zwischen den verbalen Bezeichnungen (Deskriptoren, Nondeskriptoren und Klassenbenennungen) und den Notationen (main class, optional class) erstellt. Die verbalen Bezeichnungen wurden sodann einer Reihe von Massnahmen unterzogen:

- Umwandlung gemischter Gross-/Kleinschreibung in Kleinbuchstaben (reine Grossbuchstaben-Wörter blieben als solche erhalten);
- Eliminierung aller Sonderzeichen;
- Eliminierung von Stoppwörtern auf Basis einer externen Datei;
- Eliminierung aller ein- und zweibuchstabigen Wörter, um falsche Treffer zu vermeiden;
- Eliminierung aller geographischen Namen (Klasse 950) als irrelevant für den ingenieurwissenschaftlichen Kontext;
- Umwandlung aller invertierten Thesaurusbegriffe (meist Nondeskriptoren) in Boolesche Ausdrücke;
- Stammformenbildung ("stemming") mittels des Algorithmus von Porter (1980); dies wurde jedoch nur als Option betrachtet und in der endgültigen Version nicht eingesetzt (s.u.)

Aufgrund dieser Bearbeitungsschritte resultierte ein Thesaurusformat mit Dreiergruppen ("triplets") des folgenden Inhalts:

(a) Gewicht (s.u.);

(b) Term: Einzelwort / Phrase (Mehrwortgruppe) / Boolescher Ausdruck (nur mit "@and"; da ODER-Verknüpfungen in zwei bzw. mehr Terme zerlegt wurden);

(c) Notation(en).

Dieses Vokabular wies insgesamt über 3.000 Einzelwörter und fast 18.000 zusammengesetzte Begriffe auf (Phrasen bzw. Boolesche Ausdrücke). Nach der Aufbereitung verfügte es über 13.586 Terme mit zugeordneten Hauptnotationen und 7.355 Terme mit optionalen Notationen. In Summe wurden 854 verschiedene Notationen mit durchschnittlich 25 Term-Zuordnungen verwendet; pro Hauptnotation wurden im Mittel

[15] 2. Auflage, 1995.

11 Begriffe zugeordnet. Auf diese Weise war ein reichhaltiges und weit über die reinen Klassenbenennungen hinausgehendes fachspezifisches Vokabular entstanden.

3.2.3 Klassifizierungsprozess

Abbildung 3-1 veranschaulicht schematisch die Schritte des im Projekt *DESIRE II* eingesetzten automatischen Klassifizierungsprozesses. Nach dem Harvesting eines Dokuments aus dem Web wird dessen Text extrahiert und einem Matching mit dem oben beschriebenen Vokabular unterzogen. Durch Anwendung bestimmter Regeln ("Heuristiken") kommt es zur Zuordnung der relevantesten Notation(en), wonach das Ergebnis für die Darstellung am Web bzw. die Speicherung in einer Datenbank formatiert wird.

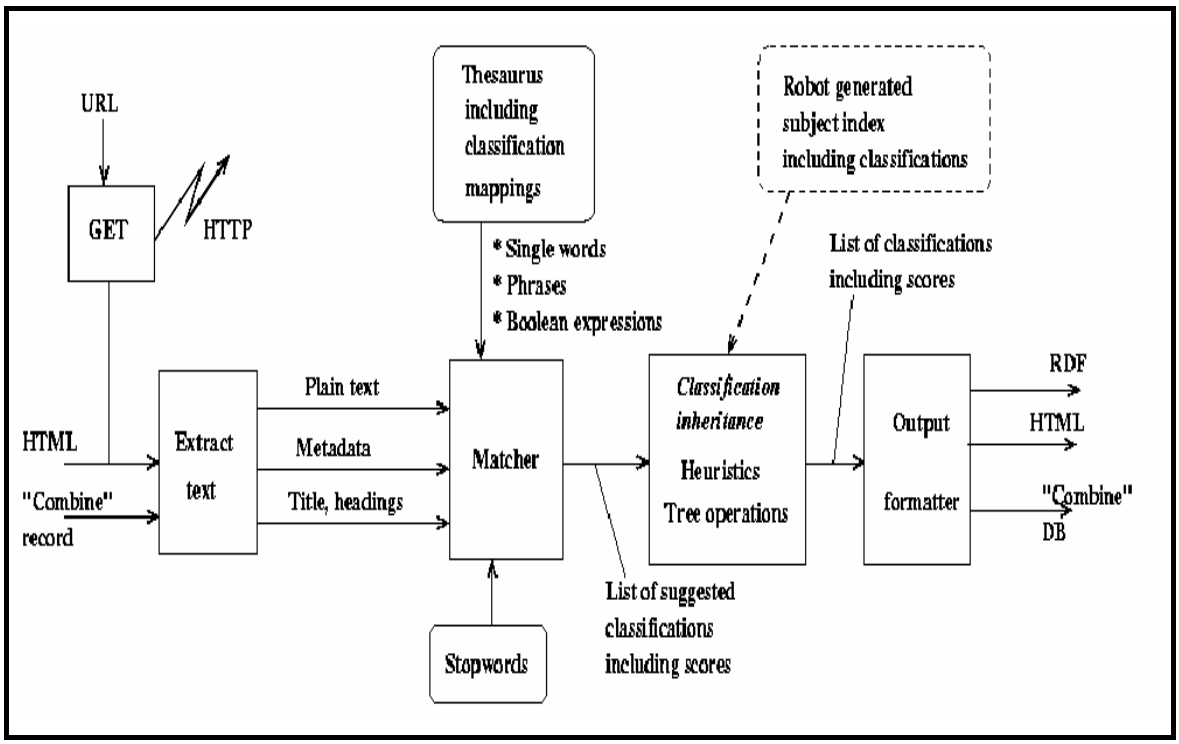

Abbildung 3-1: *DESIRE II* – Prozess des automatischen Klassifizierens (Quelle: Koch & Ardö 2000b)

Textaufbereitung. Im Zuge der Extrahierung des Textes der durch die Harvesting-Software bereitgestellten 155.611 Web-Dokumente wurden die gleichen Aufbereitungsschritte durchgeführt wie oben für den Thesaurus beschrieben (Gross-/Kleinbuchstaben, Stoppwörter usw.) Mittels eines Parsers wurde der so bereinigte Text drei Kategorien ("mark-up groups") zugeordnet, die bei der Gewichtung eine Rolle spielen sollten:

(a) Metadaten (aus den <meta>-Tags des HTML-Textes);

(b) Wichtiger Text (Titel des Dokuments; alle HTML-Headings);

(c) Normaler Text (Rest).

Mithilfe eines Spracherkennungsverfahrens wurden die nicht englischsprachigen Dokumente identifiziert und ausgeschieden (rund 15%); für die weitere Verarbeitung verblieben 132.120 Datensätze.

Matching. Für jede dieser Kategorien wurde sodann ein Matching mit dem gesamten Thesaurus-/Klassifikations-Vokabular versucht. Bei Übereinstimmung wurden dem Dokument alle mit dem betreffenden Term verknüpften Notationen als "Kandidaten" zugeordnet, zusammen mit einem Gewicht, das von mehreren Faktoren abhängig war (s.u.) Das Matching erfolgte "case-sensitive", sodass etwa Akronyme wie BASIC oder LED von Homographen in Kleinbuchstaben unterschieden werden konnten. Wenn in einem längeren Dokument derselbe Term mehrmals gefunden wurde, so zählte jedes Vorkommen. Während bei Einzelwörtern und Phrasen die exakte Übereinstimmung gewertet wurde, zählte bei Booleschen Ausdrücken das Minimum der gefundenen Übereinstimmung der einzelnen Elemente in der jeweiligen Kategorie (z.B. Headings). Intellektuell erkannte Fehlerquellen wie z.B. "tar" (Dateityp-Erweiterung bzw. Fachbegriff für Teer und ähnliche Substanzen usw.) wurden aus prinzipiellen Gründen *nicht* eliminiert.

Für etwa 11% der Dokumente (14.199) konnte *keine* Klassenzuordnung vorgenommen werden; die restlichen 117.921 Dokumente erhielten durchschnittlich je 13 Kandidaten zugeordnet.

1 term complexity and type of classification

	Single term	Boolean	Phrase
Main classif.	2	3	8
Optional classif.	1	2	4

2 match frequency

absolute frequency of matches

3 normalization of match frequency

100 / document length in location (plain text, title and headings, metadata)

4 match location

Plain text	Title and headings	Metadata
1	2	4

Formula: 1 x 2 x 3 x 4

Abbildung 3-2: *DESIRE II* – Gewichtungsalgorithmus (Quelle: Koch & Ardö 2000b)

Gewichtung. *Abbildung 3-2* zeigt den verwendeten Gewichtungsalgorithmus, der drei Kriterien multiplikativ berücksichtigte:

(a) *Komplexität des Terms und Art der Notation:* Je komplexer der Term war (Einzelwort – Boolescher Ausdruck – Phrase), desto höher war das vergebene Gewicht, zudem in Abhängigkeit davon, ob es sich bei der entsprechenden Notation um eine "main class" oder "optional class" handelte.

(b) *Kategorie ("match location"):* Je "wichtiger" der mit dem Term übereinstimmende Text war (einfacher Text, Titel/Überschriften, Metadaten), umso höher war das vergebene Gewicht.

(c) *Häufigkeit der Übereinstimmung:* Hier wurde nach einigem Experimentieren eine normalisierte, d.h. an der Länge der jeweiligen Kategorie orientierte Gewichtung vorgenommen. Die Verwendung der absoluten Termfrequenz hätte langen Dokumenten die höchsten Gewichte gegeben und viel zu viele Dokumente in jene Klassen sortiert, die mit den häufigsten Begriffen verknüpft waren. Die Verwendung der relativen Termfrequenz hätte dagegen kurze Dokumente zu stark bevorzugt.

Nach Koch & Ardö (2000b, 8–9) basierte dieser Algorithmus auf den Erfahrungen, die mit Ranking-Methoden im Zuge des bereits erwähnten "Nordic Web Index" sowie bei anderen Retrievalsystemen gewonnen worden waren, auf der Analyse der Dokumente in der Testdatenbank und auf dem Ausprobieren verschiedener Lösungen. "The rest is heuristics" (ibid.)

Am Ende des Prozesses wurden alle für ein Dokument ermittelten Gewichte pro Klasse zusammengezählt und mittels "downward propagation" angereichert; letzteres bedeutet, dass für jede spezifischere Klasse die Gewichte der in der Baumstruktur darüberliegenden Klassen hinzugefügt wurden, was praktisch einer "Vererbung" der Gewichte gleichkommt. Pro Dokument wurde sodann eine Liste der Kandidaten-Notationen in abnehmender Reihenfolge der Gewichte erstellt.

Bestimmung der endgültigen Klassen. Um die letztlich für das Browsing-System zu verwendenden Klassen festzulegen, musste diese Kandidatenlisten an einer bestimmten Stelle abgeschnitten werden ("cut-off point", "threshold"). Das Projektteam experimentierte in diesem Zusammenhang mit mehreren Optionen:

- Zuordnung der *n höchstgereihten Notationen*: Diese Möglichkeit wurde als zu vereinfachend verworfen, da nicht einzusehen sei, dass für jede Dokument dieselbe Anzahl von Klassen zutreffend sei.
- Verwendung eines *absoluten Schwellenwertes*: Bei dieser Variante werden alle Kandidaten unterhalb eines festgesetzten absoluten Gewichtes (z.B. 10) ausgeschieden.
- Verwendung eines *relativen Schwellenwertes:* Dies bewirkt, dass jene Klassen – mehrere oder auch nur eine einzige –, die einem bestimmten Anteil aller Gewichte entsprechen (z.B. 90%), beibehalten bzw. definitiv zugeordnet werden.
- Eine vierte Möglichkeit – die Berücksichtigung der Klassen nur jeweils eines oder zweier Äste des Klassifikationsbaumes – wurde als zu riskant empfunden.

Als beste Lösung wurde schliesslich eine Kombination eines relativen Schwellenwertes mit einem absoluten Schwellenwert erachtet: Alle Kandidaten mit einem Ge-

wicht von über 3% aller Gewichte für das betreffende Dokument und einem absoluten Gewicht von über 2 wurden beibehalten, d.h. den Dokumenten definitiv zugeordnet. Damit wurden aus 1,725 Millionen Kandidaten-Klassen 625.000 ausgewählt, was eine Vergabe von durchschnittlich etwa 5 Notationen pro Dokument sowie die Nutzung von zirka 90% der insgesamt vergebenen Gewichte bedeutete. Durch die gewählte Kombination (die auch als "3,2" bezeichnet wurde) gelang es, die Belegung der grössten Klassen zu reduzieren und die der kleineren auszugleichen. Immerhin waren – nahezu unabhängig von den gewählten Schwellenwerten – durch das Verfahren 750 aller Klassen des *Ei*-Schemas belegt worden. Diese Lösung wurde als die ausgewogenste bezeichnet; für ein wesentlich stärker auf Precision abzielendes Ergebnis analysierten die Autoren z.B. eine "9,5"-Kombination (relativer Schwellenwert: 9%; absoluter Schwellenwert: 5).

Stammformenbildung und Einzelwörter. Wie oben erwähnt wurde, erhielten etwa 11% der Dokumente *keine* Klassenzuordnung. Dieser Anteil bezog sich auf die Lösung ohne Einsatz des Stemming-Algorithmus; wurde dieser verwendet, so sank der Anteil der nichtklassifizierten Dokumente auf 2%. Die Stammformenbildung hätte auch insgesamt eine wesentlich höhere Zahl von Kandidaten-Notationen, gleichzeitig aber deutlich mehr falsche Zuordnungen erbracht. Letzteres war vor allem dadurch bedingt, dass die im Matching-Prozess grundsätzlich dominierenden Einzelwörter – die ohnedies leichter zu Fehlern führten als etwa Phrasen – durch das Stemming weiter verkürzt und damit noch fehleranfälliger gemacht wurden. Daher wurde letztlich auf den Einsatz des Stemming-Verfahrens verzichtet.[16]

Durch die Analyse des Matching-Verhaltens der häufigsten Einzelwortbegriffe aus dem Thesaurus wurde auch evident, dass eine Reihe davon – insbesondere allgemeine und mehrdeutige Begriffe – zu einer quantitativen Aufblähung gerade der grössten Klassen führte. Aus der Liste der hundert häufigsten Einzelwörter wurden 26 als besonders problematisch identifiziert (z.B. "control", "engineering", "research", "technology", "processing") und auf eine erweiterte Stoppwortliste gesetzt. Diese Massnahme reduzierte den Umfang der grössten Klassen beträchtlich, ohne jedoch die grundsätzliche Verteilung von grösseren und kleineren Klassen bzw. die Zahl der insgesamt vergebenen Notationen zu beeinflussen.

3.2.4 Evaluierung

Um das eingesetzte Verfahren zu evaluieren, wurden vier verschiedene Ansätze gewählt:

[16] Genauer dazu, auch unter Verwendung nur eines absoluten Schwellenwertes, vgl. Ardö & Koch (1999a, 243–244) sowie Koch & Ardö (2000b, 11–13).

Test des Algorithmus bzw. der Software. Um die korrekte Arbeitsweise des Verfahrens zu überprüfen, wurde es auf die via *EELS* angebotenen Web-Dokumente angewandt, die, wie erwähnt, eine intellektuelle Erschliessung mittels Thesaurus und Klassifikationssystem von *Ei* aufwiesen. Trotz des komplizierten Gewichtungsverfahrens und des heuristischen Zuordnungsansatzes erbrachte dieser Test eine zu 100% korrekte Zuordnung.

Analyse der Verteilungseffekte. Hier wurde der Frage nachgegangen, zu welcher Verteilung der Dokumente über die Klassen des *Ei*-Schemas das Verfahren führte. Es zeigte sich, dass sich diese Verteilung *nicht* mit der der Thesaurusbegriffe auf die Klassen deckte und dass sich die durch das Harvesting aufgefundenen Web-Dokumente auch nicht gleichmässig auf die Hauptgruppen aufteilten. Dennoch wurde eine zufriedenstellende Verteilung über das gesamte Klassifikationssystem – mit Ausnahme jener Klassen, mit denen keine Thesaurusbegriffe assoziiert waren (meist Hauptklassen) – erzielt. Nach Meinung der Autoren wäre allerdings für die relativ grosse Menge von Dokumenten eine etwas tiefer gegliederte Hierarchie von Vorteil gewesen.

Vergleich der automatischen mit der intellektuellen Klassifizierung. Zu diesem Zweck wurden in einer Stichprobe von Dokumenten, die sowohl in der intellektuell klassifizierten *EELS* als auch in der Testdatenbank enthalten waren, die jeweils vergebenen Notationen verglichen. Die Übereinstimmung des automatischen Verfahrens mit der intellektuellen Klassifizierung ("correct or finer") betrug ohne Einsatz des Stemming-Verfahrens 57% und mit Stemming 60%. Korrekt bis zu den drei ersten Stellen der Notationen waren 64% bzw. 66% (ohne bzw. mit Stemming). Die Autoren leiteten aus diesen Daten die Aussage ab, dass das allgemeine Ausmass der Übereinstimmung zwischen 57% und 66% liege (Ardö & Koch 1999a, 244; Koch & Ardö 2000b, 16).[17] Zur Relativierung dieser nicht sehr hohen Quote wurde argumentiert, dass etwa die Studie von Larson (1992) mit dem besten der dort getesteten Verfahren bloss 46% korrekte Zuordnungen erzielt habe und dass die intellektuelle Klassifizierung durchaus auch nicht immer korrekt und konsistent sei. Immerhin wurde aber auch eingeräumt, dass das automatische Verfahren das Problem des "Kontextverlusts" ("problem of lost context") aufweise, welches menschliche Klassifizierer nicht beeinträchtige.

Evaluierung durch Experten. Schliesslich bewertete auch eine Gruppe von Fachreferenten und Ingenieuren die Ergebnisse des Klassifizierungsverfahrens, wobei die in bestimmten Fachgebieten automatisch getroffenen Zuordnungen als fachlich korrekt bzw.

[17] Dies ist m. E. nicht korrekt – die Aussage müsste entweder "zwischen 57% und 60%" (korrekt oder feiner *ohne* bzw. *mit* Stemming) oder "zwischen 57% und 64%" (korrekt/feiner bzw. bis zu den ersten drei Stellen richtig, jeweils ohne Stemming) lauten.

nicht korrekt eingestuft wurden. *Tabelle 3-1* zeigt, dass dabei beträchtliche Unterschiede zutage traten:

Ei Klasse	Fachgebiet	Korrekt klassifiziert
903	Information science	75,5%
801.2	Biochemistry	61,5%
412	Concrete	37,0%

Tabelle 3-1: *DESIRE II* – Expertenurteile zu Klassifizierungsergebnissen (nach Koch & Ardö 2000b)

Im Durchschnitt betrug der Anteil der als korrekt bewerteten Zuordnungen 59% und lag damit etwa ähnlich wie der Grad der Übereinstimmung zwischen automatischer und intellektueller Klassifizierung. Als Grund für die grossen Differenzen zwischen den genannten Wissenschaftsgebieten wurde das zuwenig kontrollierte Verhalten der Experten angeführt (trotz gegenteiliger Anweisung beurteilten diese z.T. nicht nur die Korrektheit der Zuordnung, sondern auch die Brauchbarkeit und Qualität der betreffenden Dokumente). Aus der Analyse der besonders schlecht plazierten Dokumente wurden aber auch drei wichtige Erkenntnisse gewonnen:

(a) Der Hauptgrund für die Fehlklassifizierungen lag in nicht ausreichend disambiguierten Begriffen aus dem Thesaurus-Vokabular, wenn diese mit mehreren verschiedenen Fachgebieten assoziiert waren (Beispiel: "drives").

(b) Ein zweiter Grund – dem ersten nicht unähnlich – lag im Auftreten von natürlichsprachlichen Homonymen im Text der Dokumente, die im Thesaurus-Vokabular nur durch Begriffe aus einem anderen Kontext abgedeckt waren (Beispiel: "concrete").

(c) Dokumente aus Randgebieten der Ingenieurwissenschaften wurden mitunter zwar nach den *Ei*-Regeln korrekt klassifiziert, hingegen nach Meinung der Experten – wohl zurecht – nicht (Beispiel: Biochemie vs. Mikrobiologie).

Während Koch & Ardö (1999b, 19) meinten, dass die beiden ersten Probleme durch Disambiguierungsversuche mittels des Klassifikationssystems sowie durch den Einsatz linguistischer Komponenten bewältigt werden könnten, sahen sie für den dritten Fall, zumindest bei Verwendung eines fachspezifischen Vokabulars, keine reelle Lösungsmöglichkeit.

3.2.5 Benutzung

Wie bereits oben erwähnt wurde, werden sowohl das SBIG *EELS* als auch die Datenbank *AE* nicht mehr aktualisiert, doch können beide noch auf dem früheren Stand recherchiert werden. Aufgrund der Inspektion sowohl der *AE*-Homepage[18] als auch der für das "cross-searching" von *EELS* und *AE* eingerichteten "AEELS"-Seite[19] ist zu ver-

[18] http://eels.lub.lu.se/ae/index.html [01.03.2004]

[19] http://eels.lub.lu.se/aeels/search.html [15.05.2004]

muten, dass die im Arbeitsplan (Koch 1999) anvisierte und auch in anderen Dokumenten genannte Implementierung eines Pilotdienstes für das klassifikationsbasierte Browsing in *AE* bzw. das "cross-browsing" von *EELS* und *AE* (Koch & Ardö 1999b, 20; Cross et al. 2000b, 15) wohl nie zustande gekommen ist und somit wohl auch keine entsprechenden Endbenutzererfahrungen vorliegen.

Als Nebenprodukt des Klassifizierungsprojektes entstand ein über die "Demonstrationsseite" (Ardö & Koch 2000) zugänglicher WWW-Klassifizierungsdienst, bei dem Endbenutzer die Adresse eines beliebigen Web-Dokuments mit ingenieurwissenschaftlichem Bezug eingeben, Gewichte und Schwellenwerte wählen und die resultierende Klassifizierung (*Ei*-Klassen) als HTML-Seite oder als RDF-Struktur betrachten konnten. Über die Benutzung dieses Dienstes, dessen Webadresse inzwischen leider nicht mehr gültig ist, ist jedoch nichts bekannt.

3.2.6 Anwendung anderer Klassifizierungsverfahren

Aufgrund von Kooperationen mit dem BIS Oldenburg und mit OLCL (Dublin, OH) konnten die im *DESIRE*-Projekt verwendeten Dokumente zu Testzwecken auch mit zwei anderen Verfahren – *GERHARD* und *Scorpion* – automatisch klassifiziert werden. Die Ergebnisse dieser Versuche werden bei der Darstellung dieser beiden Projekte behandelt (vgl. Abschnitte 5.2 und 6.7).

3.2.7 Thematisches Vorfiltern beim Web-Harvesting

Im Zuge der Beschäftigung mit dem automatischen Klassifizieren entstand im *DESIRE*-Projektteam auch die Idee, das Verfahren zum Zweck des thematischen Vorfilterns (hier "topic filtering" bzw. "subject filtering" genannt, aber auch bekannt als "focused crawling" bzw. allgemein als "Informationsfiltern") im Harvesting-Prozess einzusetzen. Dies liegt insbesondere beim Suchen von Dokumenten für ein roboter-generiertes SBIG nahe, da auf diese Weise thematisch nicht relevante Dokumente erst gar nicht in den Klassifizierungsprozess gelangen. Nach den Vorstellungen der Autoren (Ardö et al. 1999; Koch & Ardö 2000b) würde die dazu nötige Variante ihres Harvesters COMBINE zwar eine ähnliche Übereinstimmungsprüfung wie oben beschrieben durchführen, dafür aber nur eine gewichtete Begriffsliste und kein Klassifikationssystem benötigen.[20]

3.2.8 Exkurs: SOSIG

Wie eingangs erwähnt wurde, war einer der Projektpartner bei *DESIRE II* das ILRT (Bristol), das auch Trägerinstitution des bekannten britischen SBIGs "Social Science Information Gateway" (SOSIG) ist. *SOSIG* diente, wie auch EELS und der niederländi-

[20] Die von den Autoren beschriebene Demonstrationsversion (Thema: "Fleischfressende Pflanzen") ist jedoch nicht mehr am WWW auffindbar.

sche Dienst "DutchESS",[21] bereits im Projekt *DESIRE I* zum Testen und Demonstrieren der Funktionen von SBIGs. Da in der Literatur neben *EELS* mitunter auch *SOSIG* im Zusammenhang mit dem automatischen Klassifizieren erwähnt wird, lag es nahe, der Frage nachzugehen, ob im Rahmen von *DESIRE II* auch mit diesem SBIG experimentiert wurde. Diese Recherchen erbrachten jedoch lediglich eine kurze Passage, in der erwähnt wird, dass *SOSIG* die im Projekt *DESIRE II* entwickelten Werkzeuge zum automatischen Klassifizieren adaptiert habe, um die Möglichkeiten ihres Einsatzes für den eigenen "Harvester Index" zu untersuchen (Hiom 2000, 57). Weder aus der *SOSIG*-Homepage[22] noch aus der übrigen zu diesem SBIG erschienenen Literatur (Cross 2000; Cross et al. 2000a; Hiom 1998; Maggs 1999; Monopoli & Nicholas 2000; Worsfold 1997) lässt sich jedoch der geringste Hinweis zu diesem Thema finden, sodass angenommen werden kann, dass hier keine relevanten Aktivitäten vorliegen dürften.

3.3 Engine-e

Engine-e ist die Bezeichnung für einen roboter-generierten Web-Index mit einem klassifikationsbasierten Browsing-Interface, der die ingenieurwissenschaftlichen Disziplinen abdeckt (Lindholm et al. 2003; Schönthal 2003). Das System wurde nach der aus finanziellen und anderen Gründen erfolgten Einstellung von *EELS* auf Basis der Erfahrungen der NetLab-Projekte gemeinsam mit der Bibliothek der Technischen Hochschule Stockholm entwickelt und ist – allerdings nur als Prototyp – seit Anfang 2003 öffentlich verfügbar.[23] Ziel war es, qualitative Aspekte eines SBIG (Erschliessung, Browsing-Struktur) mit den quantitativen eines roboter-generierten Web-Index (Datenmenge) zu kombinieren und dabei den hohen Personalaufwand für die sachliche Erschliessung – bei Akzeptanz einer etwas geringeren Erschliessungsqualität – zu vermeiden.

3.3.1 Methodische Vorgangsweise

In methodischer Hinsicht wurde sehr ähnlich wie im Rahmen von *DESIRE II* vorgegangen. Ein wichtiger Unterschied liegt darin, dass bei *Engine-e* die automatische Klassifizierung bereits in den Harvesting-Prozess (mit der Software COMBINE) eingebunden wurde, sodass sie nicht nur der sachlichen Erschliessung der Web-Dokumente, sondern auch dem Informationsfiltern diente. Neben HTML-Seiten wurden nunmehr auch PDF-Dokumente berücksichtigt. Der Klassifikator schied etwa 80% der durch das Harvesting eingebrachten Dokumente aus; die erfolgreich klassifizierten Dokumente wurden in der *Engine-e* Datenbank gespeichert und die in ihnen enthaltenen Links dem

[21] http://www.klb.nl/dutchess/ [02.06.2004]

[22] http://www.sosig.ac.uk/ [22.02.2004]

[23] http://engine-e.lub.lu.se/ [09.06.2004]

Roboter zur weiteren Fortsetzung des Harvesting übergeben. Auf diese Weise entstand über einen Harvesting-Zeitraum von sieben Wochen eine Sammlung von rund 350.000 automatisch klassifizierten Dokumenten.[24]

Als Klassifikation wurde wiederum das Schema von *Ei* verwendet; die Vokabularanreicherung erfolgte abermals auf der Basis des *Ei*-Thesaurus. Das Vokabular umfasste bereits fast 21.000 Begriffe (Einzelwörter, Phrasen, Boolesche Ausdrücke). Gewichtung und Zuteilung der endgültigen Notationen erfolgten im Prinzip wie bei *DESIRE II*; über die konkreten Gewichte und Schwellenwerte ist nichts bekannt, ausser dass neuerlich experimentiert wurde, um geeignete Parameter zu finden.

3.3.2 Evaluierung

Die vorliegende Literatur erwähnt keine systematische Evaluierung des Klassifizierungsverfahrens. Allerdings wird berichtet, dass im Zuge der Erstellung von *Engine-e* wiederholt den Fragen nachgegangen wurde, ob zuviel Material geringer Qualität akzeptiert wurde, hochwertige Dokumente verloren gingen, relevante Klassifizierungsergebnisse erzielt und zuviele bzw. zu wenige Notationen pro Dokument vergeben wurden. Die praktischen Erfahrungen mit dem fertigen Prototyp zeigten schliesslich, dass dieser oft einen hohen Recall erbrachte, die Precision jedoch nicht immer zufriedenstellend war.

3.3.3 Benutzung

Das Web-Interface zur Benutzung von *Engine-e* sieht sowohl eine Suchfunktion (mit mehreren Optionen) als auch eine Browsing-Struktur vor, die mit den sechs Hauptklassen des *Ei*-Schemas beginnt. Durch Anklicken einer dieser Klassen erfolgt die Anzeige der Notationen und Bezeichnungen der darunterliegenden Klassen usf., wobei jeweils auch die Anzahl der durch die jeweiligen Klassen repräsentierten Dokumente ausgegeben wird (*Abbildung 3-3*). Auf jeder Ebene besteht die Möglichkeit, zur Anzeige der Browsing-Ergebnisse zu wechseln, bei der zunächst die Titel der Dokumente (in alphabetischer Reihenfolge) angezeigt werden, deren Auswahl schliesslich zur Vollanzeige der betreffenden Web-Ressource führt. Da diese Vollanzeige auch verlinkte Schlagwörter – vermutlich die im Klassifizierungsverfahren für die endgültige Zuordnung verantwortlichen Thesaurusbegriffe – enthält, können in der Folge alle weiteren Dokumente mit dem betreffenden Schlagwort aufgerufen werden.

Im Herbst 2003 wurde durch das Projektteam eine Benutzerbefragung durchgeführt, bei der *Engine-e* positiv beurteilt wurde. Aufgrund dieses Benutzerfeedbacks sollte allerdings die alphabetische Titelanzeige durch ein Relevanz-Ranking (wie bei der

[24] Die zitierte Literatur erwähnt kurz, dass der kombinierte Prozess von Harvesting und Klassifizierung quantitativ leicht ausufern kann ("grow out of hand") und daher Zwischenkontrollen unumgänglich seien.

Suchfunktion bereits implementiert) ersetzt werden. Ausserdem wurde die Option der Limitierung von Suchen mittels der (trunkierten) Klassifikation gewünscht (Lindholm et al. 2003, 5).

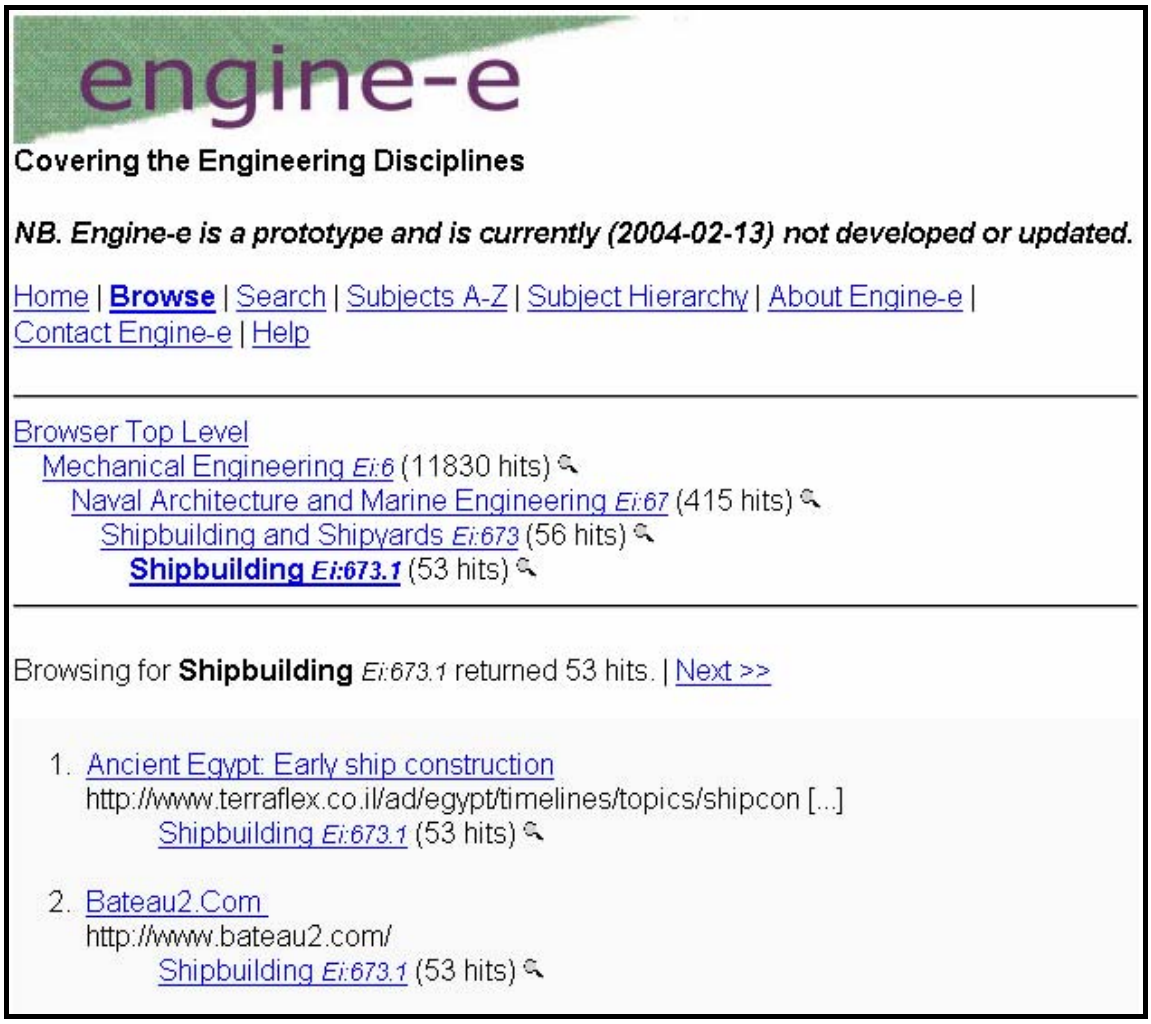

Abbildung 3-3: *Engine-e* – Benutzerinterface
(Quelle: http://engine-e.lub.lu.se/)

4 Wolverhampton Web Library (The UK Web Library)

Auch an der *School of Computing and Information Technology* der Universität Wolverhampton (Grossbritannien) begann man sich in der ersten Hälfte der 1990er Jahre für eine klassifikatorische Erschliessung des WWW zu interessieren. Im Gegensatz zu den fachlich orientierten Projekten *AE* bzw. *Engine-e* ging es hier um die Aufbereitung von nationalen, d.h. *britischen* Ressourcen (vorwiegend institutionellen Webseiten) *aller Disziplinen* – daher auch die Bezeichnung *The UK Web Library*. Das Unternehmen begann mit der Erstellung einer vollständig manuell generierten und gewarteten Linksammlung, deren ursprüngliche Gliederung alsbald Probleme aufwarf. Aufgrund bereits bestehender (bibliotheksbezogener) Erfahrungen mit der DDC entschloss man sich, dieses Klassifikationssystem für die fachliche Strukturierung der Linkliste bzw. das Browsing in derselben heranzuziehen, zumal damit ein universelles, hierarchisches und vielen Benutzern aus Bibliotheksanwendungen bereits bekanntes System vorlag. Damit war der bis heute bestehende Dienst *Wolverhampton Web Library* bzw. *WWLib* entstanden (Burden 1997).

4.1 WWLib-TOS und "Old ACE"

Die ursprünglich erstellte Software ("The Original Software" oder "TOS") sollte zunächst eine einfache Suchmöglichkeit sowie eine nach der DDC-Hierarchie gestaltete Ergebnisdarstellung ermöglichen. Dazu kamen in der Folge die Option der Recherche auch nach Dewey-Notationen sowie das Browsing in den Hierarchien der DDC; pro Web-Dokument war auch mehr als nur eine Notation möglich (Burden 1997). Die sachliche Erschliessung selbst erfolgte intellektuell – *WWLib* unterschied sich somit zu Beginn nicht von anderen kategorisierten Web-Katalogen, ausser dass eine aus dem bibliothekarischen Bereich stammende Klassifikation verwendet wurde. Mit steigender Zahl der verzeichneten Ressourcen (ca. 2.000–5.000) begann man jedoch – um 1994 – an eine Automatisierung des Klassifizierungsprozesses zu denken und entwickelte dafür eine als *ACE* ("automatic classification engine") und später als *Old ACE* bezeichnete Software. Die folgende Darstellung der Vorgangsweise beim automatischen Klassifizieren folgt vor allem den Ausführungen von Wallis & Burden (1995) und Burden (1998a).

4.1.1 Aufbereitung des DDC-Vokabulars

Für *WWLib* kam (mit Zustimmung von OCLC Forest Press) die 20. Auflage der DDC zur Anwendung. Als Voraussetzung für das automatische Klassifizieren wurde zunächst

eine als "Thesaurus"[1] bezeichnete Konkordanz zwischen den Notationen und den Klassenbenennungen geschaffen, die allerdings deutlich weniger umfangreich als die gedruckten DDC-Tafeln war, jedoch viele Änderungen von Details[2] sowie fallweise die britische anstelle der amerikanischen Schreibweise aufwies. Das die Klassen repräsentierende Vokabular wurde mit einem neu entwickelten Stemming-Algorithmus behandelt, der im Gegensatz zu dem im *DESIRE*-Projekt getesteten Verfahren von Porter (1980) darauf abzielte, *ganze* englische Wörter mit verschiedenen Endungen ("extended derivatives") zu bilden, also z.B. aus Wörtern mit der Endung "-graphy" Wörter mit Suffixen wie "-graph", "-grapher", "-graphic" und "-graphically" abzuleiten; daneben führte das Verfahren auch eine Depluralisation und Degerundisation durch (Burden 2000).[3] Die interne Speicherform des "Thesaurus" inkludierte auch die Position der einzelnen Wörter in den jeweiligen Klassenbenennungen, wobei die durch das Stemming-Verfahren gebildeten Synonyme jeweils ein und derselben Position zugeordnet wurden.

4.1.2 Klassifizierungsprozess

Allgemein formuliert, wurde auch durch dieses Verfahren ein Wort-für-Wort-Vergleich zwischen dem Text der Dokumente und dem DDC-Vokabular durchgeführt. Beim Auftreten einer Übereinstimmung wurden die beiden Wortsequenzen (Dokument und DDC) "synchronisiert", um durch die Analyse des davorliegenden Textes eine mögliche (partielle oder vollständige) Übereinstimmung von aus dem DDC-Vokabular entnommenen Phrasen zu entdecken; beim Auftreten jedes "match" wurde die Punktezahl für die betreffende DDC-Klasse erhöht. Nach dem Ende des Prozesses erfolgte pro Web-Ressource die Ermittlung bzw. Zuordnung der Klasse mit dem höchsten Punktestand.

Textaufbereitung. *Old ACE* war nur für HTML-Dokumente entwickelt worden. Der Text jedes dem Klassifizierungsverfahren zugeführten Web-Dokuments wurde im Zuge des "input parsing" mehreren Aufbereitungsschritten unterworfen:

(a) Der grösste Teil der HTML-Codierung wurde ignoriert. Eine Ausnahme bildeten die Tags für Titel ("<title>") und Hauptüberschriften ("<H1>", "<H2>"), da der hievon umschlossene Text als bedeutender betrachtet und daher mit einem höheren Gewicht ausgestattet wurde als die übrigen Wörter. Das <meta>-Tag wurde dagegen ignoriert, da negative Auswirkungen der bekannten Praxis, dieses zur Beeinflussung des Rankingverhaltens von Suchmaschinen zu nutzen, befürchtet wurden.[4]

---

[1] Die etwas unglückliche Verwendung dieses Terminus' wurde später auch zugestanden: "This file is called the *thesaurus* for historical reasons although, strictly, this is a bit of a misnomer." (Burden 1998a, 1).

[2] So musste etwa die in dieser DDC-Ausgabe nicht adäquate Strukturierung des Fachgebietes "computer science" verbessert werden.

[3] "depluralisation" = Bildung der Einzahlform aus Wörtern mit Endungen wie "s", "es", "ies"; "degerunding" = Bildung der äquivalenten Verbform aus Wörtern mit der Endung "ing".

[4] Später kam man aber zu der Ansicht, dass dieser wie auch der im "alt"-Attribut von <img>-Tags enthaltene Text durchaus nützlich sein könnte (Burden 1998a, 2).

(b) Punkte, Kommas usw. wurden bei der Worterkennung eliminiert, doch kam ihnen für die Bestimmung der Wortposition Bedeutung zu, da beim Auftreten dieser Zeichen eine (einstellbare) Zahl von "unsichtbaren" Dummy-Wörtern erzeugt wurde, was die Wahrscheinlichkeit reduzieren sollte, dass ein Wortpaar Übereinstimmung über die Grenzen von Phrasen oder Sätzen hinweg erzielen könnte. Dies galt auch für HTML-Tags wie <p> (Absatz) und die erwähnten Tags für Titel und Hauptüberschriften ("tag break sensitivity").

(c) Wörter aus einer Stoppwortliste sowie solche, die aus nur 1–2 Buchstaben, Zahlen und nicht-alphabetischen Zeichen bestanden, wurden ignoriert. Darüberhinaus kam auch ein als "site sensitivity" bezeichneter Mechanismus zum Einsatz, der bewirkte, dass bei den Webseiten bestimmter Institutionen besonders häufig auftretende Wörter ignoriert wurden (z.B. "university" und "cambridge" für die Domäne "cam.ac.uk").[5]

(d) Der im Falle des "Thesaurus" verwendete Stemming-Algorithmus wurde für den (völlig unkontrollierbaren) Text der Webseiten als zu riskant betrachtet. Daher wurde als einziges Stemming-Verfahren die Depluralisation eingesetzt.

Matching und Gewichtung. Nach der Identifizierung eines Wortes aus dem "input stream" eines Web-Dokuments prüfte *Old ACE* zunächst, ob dieses im Vokabular des "Thesaurus" vorkam.[6] Bei Auftreten einer Übereinstimmung erfolgte sodann die Zuweisung eines aus der inversen Zahl der Klassen, in denen das betreffende Wort vorkam, gebildeten Gewichtes. Dieses Gewicht wurde bei Vorliegen eines HTML-Tags für Titel oder Hauptüberschriften noch mit einem weiteren Faktor multipliziert.[7]

Regel	Distanz	Berechnung des Gewichts
1	Both words adjacent	Multiply score by factor ***f1***
2	Same distance from synchronisation point in input stream and thesaurus class record	Multiply score by ***f2*** divided by the word distance from the synchronisation point
3	Different distances from synchronisation point	Multiply score by ***f3*** divided by the sum of the two word distances from the synchronisation point

Tabelle 4-1: *Old ACE* – Regeln für die Phrasengewichtung (Quelle: Burden 1998a)

Danach nahm das Programm eine "Synchronisierung" der Wortfolge des Input-Textes aus dem Dokument und der Wortfolge des Datensatzes der betreffenden Klasse aus dem "Thesaurus" vor. Dies diente zur folgenden, rückwärts verlaufenden Überprüfung des auf 10 Wörter dimensionierten "running buffer" des Web-Dokumentes, mittels derer nach einer möglichen Übereinstimmung von Wortpaaren (Phrasen) zwischen Dokument und "Thesaurus" – eventuell mit einer unterschiedlichem Zahl von dazwischenliegenden Wörtern – gesucht wurde. Je nach Genauigkeit einer solchen Übereinstim-

[5] Die dazu nötigen Informationen wurden aus dem Domänen-Nameserver entnommen.

[6] War dies nicht der Fall, so erfolgte lediglich die Speicherung des betreffenden Wortes in einem Logfile.

[7] Über die drei konkret als "HTML context sensitivity factors" verwendeten Gewichte ist nichts bekannt. In der Literatur wird lediglich erwähnt, dass diese einstellbar waren und hinsichtlich der optimalen Werte noch immer "subject for further investigation" seien (Burden 1998a, 5).

mung wurde das bisher ermittelte Gewicht mittels der in *Tabelle 4-1* spezifizierten Regeln weiter erhöht.[8] Ein Beispiel anhand einer Textzeile und einer Dewey-Klasse findet sich bei Burden (1998a, 6).

Bestimmung der endgültigen Klassen. Nach der Verarbeitung des gesamten Textes einer Web-Ressource wurde einfach jene Klasse ermittelt und zugeordnet, die das höchste Gesamtgewicht erzielt hatte. Anfängliche Versuche, dabei auch eine Prozedur einzusetzen, bei der die Gesamtgewichte einzelner Klassen auch den jeweils untergeordneten Klassen zugeschlagen wurden ("trickledown"), erzielten nur unbefriedigende Resultate und wurden wieder aufgegeben. Ohne dass dies in der Literatur explizit diskutiert wurde, geht daraus hervor, dass jedes Dokument offensichtlich nur *eine* Notation zugewiesen erhielt. Wenn das Gesamtgewicht zweier verschiedener Klassen gleich oder sehr ähnlich war, wurde mittels eines clusteranalytischen Verfahrens eruiert, welche dieser beiden Klassen die grössere Zahl ähnlicher Nachbarklassen aufwies und dieser sodann der Vorzug gegeben. In zwei Fällen ("discrimination") wurde hingegen, um Fehlklassifizierungen zu vermeiden, *gar keine* Notation vergeben:

(a) wenn das für eine Web-Ressource resultierende höchste Gesamtgewicht einer Klasse unter einem Wert *df1* lag;

(b) wenn das Verhältnis zwischen dem höchsten Klassengewicht und der Zahl von Klassen, die ein Gewicht von mehr als 1,5 erzielt hatten, über einem Wert *df2* zu liegen kam (dies wurde als Indikator dafür gewertet, dass die Web-Ressource eine Vielzahl von Themen aufwies und keinem davon eindeutig zugeordnet werden konnte).[9]

Für eine spätere Version von *Old ACE* wurde auch eine Verwendung der nicht im "Thesaurus" gefundenen Wörter geplant oder zumindest angedacht. Dabei wären pro Dokument die häufigsten "unbekannten" Wörter ermittelt und diese, gemeinsam mit einem noch zu bestimmenden Gewicht, der Klassenbeschreibung im "Thesaurus" zugeordnet worden. So wäre ein Mechanismus entstanden, mit dem das Programm neues Vokabular (insbesondere Eigennamen) hätte "lernen" können.

4.1.3 Evaluierung

Wallis & Burden (1995) berichten von einem ersten Experiment, bei dem *Old ACE* mit 258 bereits manuell klassifizierten Dokumenten getestet wurde. Die Ergebnisse wurden als "exact", "close" (z.B. 781 für ein intellektuell als 781.2 klassifiziertes Dokument) bzw. "none" (unklassifizierbar) kategorisiert. Dabei vermochte das Verfahren aber nur 20–22% der Web-Ressourcen korrekt ("exact" bzw. "close") zu klassifizieren; selbst

[8] Auch über die drei konkret für f1, f2 und f3 verwendeten Gewichte ist nichts bekannt. In der Literatur wird lediglich erwähnt, dass diese einstellbar waren und hinsichtlich der optimalen Werte noch immer "subject for further investigation" seien (ibid., 6).

[9] Burden (1998a, 9) gibt als Werte für df1=10 und df2=3 an, betont aber, dass diese Parameter einstellbar und noch "subject for further investigation" seien.

unter Einrechnung der unklassifizierbaren Dokumente betrug dieser Anteil nur etwa ein Drittel (33–34%).

Jenkins et al. (1997, 6) erwähnen kurz eine von *Old ACE* auf der Basis von 100 bereits vorher klassifizierten Dokumenten erzielte Klassifizierungsgüte von 40%.

Burden (1998a; 1998b) führte später ein weiteres Experiment mit 34 Web-Ressourcen durch und registrierte, dass durch die Einführung des oben erwähnten "discrimination"-Algorithmus die Erfolgsquote von 42% auf 70% gehoben wurde, da dadurch offenbar nur unrichtig klassifizierte Dokumente als nicht klassifizierbar ausgeschieden wurden. Die ausgeschiedenen Seiten wurden sodann genau analysiert, um möglichen Fehlerquellen und Verbesserungsmöglichkeiten auf die Spur zu kommen. Dabei wurde eine ganze Reihe potentieller Massnahmen entdeckt bzw. vorgeschlagen:

(a) Berücksichtigung und Gewichtung der in <meta>-Tags enthaltenen Schlagwörter und Beschreibungen;

(b) Erweiterung des "Thesaurus" (die Analyse zeigte, dass das Vokabular zu beschränkt war);

(c) Berücksichtigung von Text in Javascript-Anweisungen und im <img>-Tag ("alt"-Attribut);

(d) Untersuchung des Effekts von Stichwörtern, die im Vokabular einer Klasse mehrfach auftreten (diese wurden bislang mehrfach gewertet, was wahrscheinlich negative Effekte hatte);

(e) Untersuchung der Möglichkeit, bisher ausgeschiedene Stoppwörter dann zu berücksichtigen, wenn sie in Phrasen auftreten (z.B. "Vitamin A");

(f) Entwicklung von Mechanismen zur Erkennung von Eigennamen;

(g) Untersuchung eines Algorithmus zur höheren Gewichtung von übereinstimmenden Phrasen;

(h) Ausschluss nicht-britischer Webseiten bis zur Implementierung einer Synonymliste (entgegen der ursprünglichen Intention enthielt *WWLib* bald auch Seiten der nicht eindeutig britischen Domänen ".com", ".net" und ".org");

(i) Verbesserung der Worterkennungsprozedur.

4.1.4 Benutzung

Über Einsatz und Benutzung von *WWLib-TOS* ist nicht viel bekannt. Nach Wallis und Burden (1995) wurde *Old ACE* zunächst zur automatischen Vor-Klassifizierung der in *WWLib* neu aufgenommenen Web-Ressourcen verwendet; d.h. dass die von der Software vorgeschlagenen Notationen nur zur Unterstützung eines menschlichen Klassifizierers dienten. Eine im März 2004 vorgenommene Inspektion der *WWLib*-Homepage[10] erbrachte, dass dort offensichtlich immer noch *WWLib-TOS* und nicht die im nächsten Abschnitt beschriebene nächste Generation der Software in Betrieb ist. Aus der Seite geht auch nicht hervor, wieviele Web-Ressourcen von *WWLib* erschlossen werden bzw. wie oft dieser Dienst aktualisiert wird. Über Endbenutzerakzeptanz und/oder -verhalten konnte ebenfalls nichts eruiert werden.

[10] http://www.scit.wlv.uk/wwlib/ [01.03.2004]

4.2 WWLib-TNG und ACE

Während *WWLib-TOS* im wesentlichen auf die (Privat-)Initiative einzelner Angehöriger der *School of Computing and Information Technology* zurückging und in weiten Teilen auf manuellen Arbeitsschritten beruhte, sollte ein 1997 begonnenes offizielles Forschungs- und Entwicklungsprojekt eine zwar auf *WWLib-TOS* aufbauende, aber neu und durchgehend automatisierte Generation dieses Dienstes schaffen ("The Next Generation" oder "TNG"). Das dafür erstellte Konzept sah eine verteilte Architektur mit sechs automatisierten Komponenten vor *(Abbildung 4-1)*:

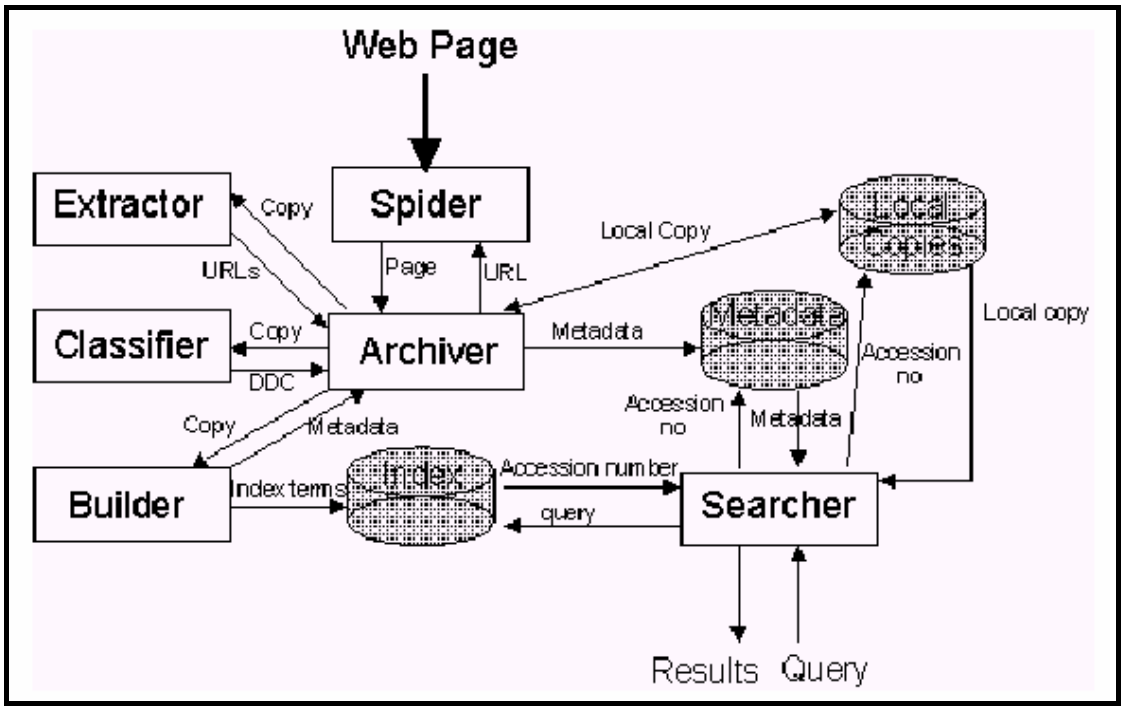

Abbildung 4-1: *WWLib-TNG* – Architektur
(Quelle: Jenkins et al. 1998)

(a) ***Spider*** – automatisches Harvesting aus dem WWW;

(b) ***Indexer*** – erhält Webseiten vom Spider, generiert ID-Nummer und Metadaten-Schablone; verteilt Kopien an Analyser, Classifier und Builder und fügt die von den beiden letzteren erzeugten Metadaten in die Schablone ein;

(c) ***Analyser*** – untersucht die erhaltenen Seiten auf Hyperlinks und übergibt gefundene URLs an Indexer, der sie vor Weitergabe an den Spider auf UK-Relevanz überprüft;

(d) ***Classifier*** – analysiert die erhaltenen Seiten und generiert DDC-Notationen;

(e) ***Builder*** – analysiert die erhaltenen Seiten und generiert Metadaten; baut Index für suchbare Datenbank auf;

(f) ***Searcher*** – Benutzerinterface.

Die Funktionsweise von mehreren dieser Komponenten wird in der Literatur näher erläutert (vgl. Burden 1997; Burden & Jackson 1999; Garratt et al. 1999). Die fol-

gende Darstellung konzentriert sich auf die Komponente "Classifier" und beruht vor allem auf den Beiträgen von Jenkins et al. (1997; 1998).

Grundsätzlich wurde festgelegt, dass der neue Klassifikator kein maschinell lernendes System sein, sondern auf manuell erstellten Repräsentationen der Klassen beruhen und von der hierarchischen Struktur der DDC Gebrauch machen sollte. Dies bedeutete, dass im Gegensatz zu *WWLib-TOS* für *WWLib-TNG*

- ein wesentlich detaillierterer "Thesaurus" mit ausführlichen Listen von Termen und Synonymen für jede Klasse erforderlich war; und dass
- das Verfahren nicht alle möglichen Klassen untersuchen, sondern bei den zehn "top classes" der DDC beginnen und nur (rekursiv) mit den Subklassen jener Hauptklassen fortsetzen sollte, bei denen eine signifikante Ähnlichkeit mit dem Dokument festgestellt würde.

4.2.1 Klassifizierungsprozess

Die für *WWLib-TNG* entworfene Klassifizierungs-"Maschine" *ACE* basiert auf einem objektorientierten Design. *Abbildung 4-2* veranschaulicht die wesentlichen Objekte des Klassifizierungsprozesses und die Beziehungen zwischen ihnen.

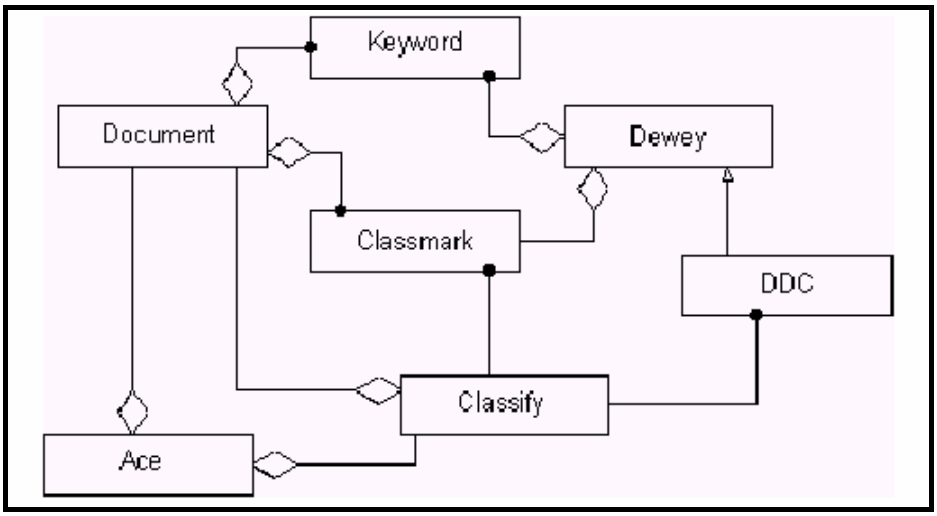

Abbildung 4-2: *WWLib-TNG* – Modell des Klassifizierungsprozesses (Quelle: Jenkins et al. 1997)

Das durch den Klassifikator (das *ACE*-Objekt) zunächst erzeugte *Document*-Objekt besteht aus einer Reihe von *Keyword*-Objekten, die jeweils ein im Dokument gefundenes Wort repräsentieren.[11] Jedem Keyword ist ein auf der Herkunft im Dokument basierendes Gewicht sowie die Wortposition im Dokument zugeordnet. Der Klassifikator verwendet ein *Classify*-Objekt, welches das neu erzeugte *Document*-Objekt mit einer Reihe von *DDC*-Objekten vergleicht. Die letzteren verfügen über die Struktur und die Eigenschaften einer abstrakten Klasse *Dewey*, die ein Objekt definiert, das eine Liste

[11] Ob bei diesem Verfahren neben Einzelwörtern auch Phrasen geprüft wurden, ist nicht bekannt, da die Aufbereitung des Klassenvokabulars ("Thesaurus") in der Literatur nicht näher ausgeführt wird.

von Stichwörtern ("keywords"), eine Liste von Subklassen sowie eine Notation ("classmark") speichert und verwaltet. Auch die *DDC*-Objekte bestehen daher aus einer Reihe von gewichteten *Keyword*-Objekten, die in ihrer Summe die jeweilige Klasse repräsentieren. Jedes *DDC*-Objekt verfügt über ein *Classmark*-Objekt (die Dewey-Notation) und kann bis zu zehn Subklassen aufweisen, die wiederum *DDC*-Objekte sind und die nächste darunterliegende Hierarchieebene repräsentieren. Das *Classify*-Objekt beginnt den Vergleichsprozess bei den zehn *DDC*-Objekten, die oberste Hierarchieebene darstellen. Wenn das Dokument signifikant mit einem *DDC*-Objekt übereinstimmt, setzt das *Classify*-Objekt mit den zugehörigen Unterklassen fort. Dieser Prozess wird rekursiv durch die Ebenen der Hierarchie weitergeführt, bis ein "leaf node" (einer Klasse ohne weitere Unterklassen) erreicht wird. Weist diese Klasse eine signifikante Übereinstimmung mit dem Dokument auf, so wird das zugehörige *Classmark*-Objekt in das *Document*-Objekt kopiert.[12] Auf diese Weise können pro Dokument auch mehrere Klassen zugeordnet werden.

Das *ACE*-Objekt verbindet das ganze System. Es akzeptiert Anwendereingaben, generiert (auf der Basis von URL oder Dateinamen) ein neues *Document*-Objekt, erzeugt für dessen Verarbeitung ein neues *Classify*-Objekt und gibt die resultierenden Notationen aus.

Der Klassifikator wurde in der Programmiersprache Java implementiert. Bei der konkreten Anwendung des Verfahrens wurden beim Matching die Gewichte der übereinstimmenden Stichwörter aus Klassenbeschreibungen und Dokumenten addiert, pro Klasse summiert und durch Multiplikation mit dem Faktor 2 und Division durch die Summe der Wortanzahl von Dokument und Klassenvokabular auf einen Bereich zwischen 0 und 1 normalisiert.[13] Resultierende Werte von über 0,5 wurden als signifikant und damit als das Kriterium für die Fortsetzung des Vergleichsprozesses gewertet.

In der Folge wurde das Verfahren auch für die automatische Erstellung von RFD-Metadaten vorgeschlagen (Jenkins et al. 1999).

4.2.2 Evaluierung

In der durch die zitierte Literatur dokumentierten Projektphase wurden lediglich die obersten 100 Dewey-Klassen verwendet, da das Erstellen des Klassenvokabulars im angestrebten Umfang ein sehr arbeitsaufwendiges Unterfangen war. Ein erster Test mit 17 zufällig ausgewählten Web-Ressourcen erbrachte 13 richtige und zwei falsche Zuord-

[12] Die Bemerkung bei Jenkins et al. (1997, 13), dass andernfalls keine Zuordnung erfolge, ist m. E. unrichtig (oder ein logischer Fehler), da in diesem Fall korrekterweise die übergeordnete Klasse, die ja noch über eine signifikante Übereinstimmung verfügen muss, zugeordnet werden sollte.

[13] Dieser Algorithmus basiert auf "Dice's coefficient"; vgl. Jenkins et al. (1997, 12–13) bzw. Jenkins et al. (1999, 1307).

nungen sowie zwei Fälle, in denen keine passende Klasse ermittelt werden konnte (Jenkins et al. 1997, 15–16).

In einer rezenteren Arbeit, in der das Verfahren nur als eine der Komponenten des Gesamtkonzepts von *WWLib-TNG* behandelt wird, findet sich die kurze Erwähnung experimenteller Ergebnisse, nach denen *ACE* 60% der klassifizierten Webseiten korrekt zugeordnet habe (Burden & Jackson 1999, 3).

4.3 Weitere Experimente mit ACE

4.3.1 Adaptives automatisches Klassifizieren mit *ACE*

In einer neueren Arbeit diskutierten Jenkins & Inman (2000) das bislang in *WWLib* praktizierte manuelle Definieren des Klassenvokabulars sowie die Vorteile einer automatischen Erzeugung von Klassenrepräsentationen auf Basis bereits klassifizierter Trainingsdokumente.

Als *Nachteile* der manuellen Generierung wurden genannt:

- Arbeitsaufwand und Langsamkeit, Notwendigkeit der Beiziehung von Fachexperten;
- Entstehen eines statischen, unflexiblen, schwierig zu aktualisierenden Vokabulars;
- Entstehen eines meist einsprachigen Vokabulars ohne einfache Übersetzungsmöglichkeit.

Vorteile der automatischen Erzeugung sah man dagegen in folgenden Aspekten:

- Entstehen dynamischer, entwicklungsfähiger, flexibler Vokabularien;
- Möglichkeit des schrittweisen Lernens durch Modifikation des Klassenvokabulars beim Hinzufügen neuer Dokumente;
- Nutzungsmöglichkeit der HTML-Struktur bei der Identifizierung wichtiger Begriffe (aus Titeln und Headings);
- Möglichkeit der Verwendung von Trainingsdokumenten in verschiedenen Sprachen zum Aufbau mehrsprachiger Klassenvokabularien;
- Wiederverwendbarkeit der Software bei Verwendung anderer Klassifikationssysteme.

Sohin wurde für *ACE* ein Vokabulargenerator programmiert, der für jede Klasse eines gegebenen Systems das diese repräsentierende Vokabular automatisch auf der Grundlage zuvor klassifizierter Dokumente erstellte. Dies geschah durch die Extrahierung der als wesentlich definierten Terme aus der Menge der jeweiligen Trainingsdokumente, d.h.:

(a) die häufig auftretenden Wörter;
(b) die in der grössten Zahl von Dokumenten auftretenden Wörter;
(c) die im Titel bzw. in Headings auftretenden Wörter.

Nach der Analyse der Trainingsdokumente wurde das vorläufig endgültige Klassenvokabular generiert, indem zunächst alle Wörter selektiert wurden, die in mehr als 10% der Trainingsdokumente aufgetreten waren; von diesen qualifizierten sich dann jene, die nach einem Gewichtungsprozess zu den höchstgereihten 50% zählten. Als Gewicht

wurde das Produkt aus der Worthäufigkeit einerseits und der Summe aus der Dokumenthäufigkeit und der Häufigkeit des Auftretens in Titeln bzw. Headings andererseits verwendet. Mit diesem Vokabular erfolgte sodann die interne Abbildung des Klassifikationssystems bzw. dessen Hierarchie. Der Prozess des automatischen Klassifizierens selbst folgte dann der oben beschriebenen Arbeitsweise von *ACE*.

Ein Test dieses Ansatzes wurde nicht mit der DDC, sondern mit der Klasse "Arts / Design_Arts / Architecture" aus dem Schema des Web-Kataloges *Yahoo!*[14] vorgenommen. Bei der automatischen Erstellung des Vokabulars wurden – soweit vorhanden – pro Klasse etwa 20 Trainingsdokumente verwendet. Die Evaluierung erfolgte mit jeweils zwei zufällig aus den 14 "leaf nodes" des Zweiges "Architecture" ausgewählten Dokumenten – in Summe also 28 Web-Ressourcen. Jeder Klassifizierungsvorgang wurde nach der Zahl der korrekt vorgenommenen Verzweigungen bewertet, was zu einem Ergebnis von 61 von 80 möglichen Punkten führte (76,2% Genauigkeit). Aufgrund der Angaben bei Jenkins & Inman (2000, 509) besteht allerdings Grund für die Vermutung, dass dieses Experiment mit Dokumenten durchgeführt wurde, die zuvor auch schon für das Training des Klassifikators gedient hatten, was es als methodisch fragwürdig erscheinen liesse (vgl. Abschnitt 2.3.1).

4.3.2 Ontologie-basiertes automatisches Klassifizieren mit *ACE*

Die neuesten Versuche zur Verbesserung der Klassifizierungsleistung von *ACE* finden sich in zwei Vorträgen von Prabowo et al. (2002a; 2002b). Aus diesen wird ersichtlich, dass hiefür eine völlig neue Version von *ACE* erarbeitet wurde, die auf der Verwendung eines semantischen Netzwerks zur Repräsentation von "domain ontologies"[15] und einem neuronalen Netz beruht, das zur Repräsentation der Relationen zwischen abstrakteren Begriffsebenen, gemeinsamen Klassen (von DDC und LCC)[16] und konkretem Klassenvokabular dient und den eigentlichen Klassifizierungsprozess durchführt. Die von *ACE* aufgebaute Begriffsdatenbank ist bestrebt, einerseits Begriffe zu definieren, die spezifisch genug sind, um von einander unterschiedenen werden zu können, und andererseits auch Begriffe, die breit genug sind, um die spezifischeren zu kategorisieren. Das System basiert auf vier strategischen Zielsetzungen zur Verbesserung der Genauigkeit der Klassifizierung:

(a) Zusammenfassung der vergebenen Gewichte auf den höheren begrifflichen Ebenen, um das "Thema" der Web-Ressource anstelle der Details herauszufinden;

[14] http://www.yahoo.com/ [16.05.2004]

[15] Hier definiert als "a single entity which holds concepts of a domain and differentiates itself from another"; "domain" wird dabei als das Fachgebiet ("subject area") verstanden, zu dem die Ontologie gehört, z.B. Informatik, Politik, Naturwissenschaften (Prabowo et al. 2002b, 2).

[16] Prabowo et al. (2002a) schlagen voll formalisierte Definitionen für das vollständige bzw. teilweise Mapping dieser beiden Klassifikationssysteme vor.

(b) Verwendung von Begriffseigenschaften und -facetten, um "Themen" zu identifizieren (auch wenn diese selbst als Begriff nicht im Dokument aufscheinen);

(c) Berücksichtigung von Phrasen und Präpositionen – im Gegensatz zu früher wird etwa "of" nicht als Stoppwort betrachtet, sondern als wichtiges Hilfsmittel zur Identifizierung von Begriffen (z.B. "anatomy of plant");

(d) Verzicht auf Depluralisation und sämtliche sonstige Stemming-Verfahren, um der Gefahr einer Bedeutungsveränderung auszuweichen.

Ein für zwei "domain ontologies" – "social sciences" und "natural sciences" – vorgenommenes Experiment zeigte eine im Vergleich zu früheren *ACE*-Versionen verbesserte Genauigkeit: Diese wurde allerdings stark zu Lasten der Komplettheit erzielt, da aufgrund der Unvollständigkeit der verwendeten Ontologien zahlreiche Dokumente gar nicht klassifiziert werden konnten.

5 German Harvest Automated Retrieval and Directory

GERHARD (German Harvest Automated Retrieval and Directory) ist die Bezeichnung für einen Such- und Navigationsdienst, der mit einem Roboter gesammelte Web-Ressourcen mit linguistischen und statistischen Verfahren automatisch klassifiziert.[1] Dieser Dienst wurde 1996–1998 im Rahmen eines DFG-Projektes durch drei Projektpartner mit der folgenden Aufgabenteilung realisiert:

- BIS (Bibliotheks- und Informationssystem der Universität Oldenburg): Leitung, Koordination, Gathering-System und dessen Steuerung;
- ISIV (Institut für Semantische Informationsverarbeitung, Universität Osnabrück): Erstellung des Klassifizierungsverfahrens;
- OFFIS (Oldenburger Forschungs- und Entwicklungsinstitut für Informatikwerkzeuge und -systeme): Datenbank und Benutzerschnittstelle.

5.1 Das DFG-Projekt GERHARD

Die Zielsetzungen des Projekts, die es auch von den damals üblichen Suchmaschinen und Webkatalogen abgrenzen sollten, sahen folgende Aspekte vor:

- geographische Beschränkung auf das *deutsche* Web-Angebot (dort aber aktueller und kompletter);
- Beschränkung auf den *Wissenschafts*bereich (Ausschluss populärer Web-Ressourcen);[2]
- Verwendung eines professionellen universellen Klassifikationssystems und ausschliesslich automatischer Verfahren, dadurch Erschliessung grösserer Mengen als bei manuell erstellten Verzeichnisdiensten;
- Integration der gezielten Suche ("searching") mit der thematischen Navigation ("browsing");[3]
- Dauerbetrieb mit geringem Ressourcenaufwand.

Die Auswahl relevanter Hosts (Universitäten, Fachhochschulen, Forschungseinrichtungen usw.) und die Ermittlung ihrer Web-Adressen erfolgten allerdings nicht vollautomatisch, sondern manuell auf Basis vorliegender Listen und Verzeichnisse. Die dabei resultierenden 350–400 Startadressen wurden mittels des Gathering-Roboters aus

[1] Die Literatur zu GERHARD ist umfangreich und repetitiv zugleich, da die meisten Publikationen zu diesem Projekt Variationen desselben Ursprungsdokuments zu sein scheinen. Die folgende Darstellung folgt, soweit nicht anders erwähnt, den Publikationen von Möller et al. (1999a; 1999b; 2000), Wätjen (1998a; 1998b) und Wätjen et al. (1998).

[2] Ursprünglich war zwar angestrebt worden, das "gesamte deutschsprachige Internet" zu indexieren, doch wurde angesichts des grossen Anteils "irrelevanter und populärer Dokumente" eine Eingrenzung auf den "wissenschaftlich relevanten Teil" beschlossen.

[3] In der Literatur zu *GERHARD* wird "browsing" häufig verkürzt mit "Navigation" gleichgesetzt; gemeint ist aber wohl stets ein thematisches (klassifikationsbasiertes) Navigieren.

dem frei verfügbaren Programmsystem "Harvest"[4] in einer Sammeltiefe von 12 Stufen abgearbeitet, wodurch innerhalb eines Jahres etwa eine Million Dokumente gesammelt wurde. Dabei erfolgte eine Beschränkung auf HTML-Dokumente sowie auf Zugriffe über das HTTP-Protokoll.

5.1.1 UDK und UDK-Lexikon

Das zu verwendende Klassifikationssystem sollte "mächtig", universell, hierarchisch, maschinenlesbar und mehrsprachig (deutsch, englisch) verfügbar sein. Zum damaligen Zeitpunkt erfüllte lediglich die UDK alle diese Kriterien; insbesondere wurde sie dem Projekt in der seit rund 20 Jahren von der ETH Zürich gepflegten maschinenlesbaren Version[5] in drei Sprachen – deutsch, englisch, französisch – und mit einem Umfang von rund 60.000 bis 70.000 Klassen[6] zur Verfügung gestellt. Die nicht ganz vollständige französische Sprachversion fand nur als Option für die Benutzeroberfläche Verwendung, während die deutsch- und englischsprachigen Klassenbenennungen daneben auch für das Klassifizierungsverfahren herangezogen wurden.

Wie eingangs erwähnt, wurden bei *GERHARD* für die automatische Klassifizierung linguistische und statistische Verfahren verwendet. Ziel der linguistischen Komponente war die Abbildung natürlichsprachlicher Phrasen, die in den HTML-Dokumenten auftraten, auf das Vokabular der UDK. Als Voraussetzung dafür war es nötig, das letztere in die Form natürlichsprachlicher Ausdrücke, die in Texten vorkommen können, zu transformieren und diese den Notationen zuzuordnen. Das Endprodukt dieses relativ aufwendigen Umwandlungsvorgangs wurde als "UDK-Lexikon" bezeichnet.

Erste Voraussetzung dafür war eine Aufbereitung der UDK hinsichtlich einer Reihe von Kriterien wie einheitliche Kleinschreibung, Normierung der Umlaute, Entfernung aller diakritischer Zeichen sowie der in den Tafeln enthaltenen Verweisungen, Anmerkungen, Kommentare und Klammerungen. Zahlreiche Einträge mussten umgeformt werden; dies betraf etwa die Auflösung invertierter Phrasen oder die Isolierung von Begriffen aus Aufzählungen.

Ein weiterer Schritt war die Anwendung linguistischer Software[7] zur morphologischen Analyse jedes Wortes der UDK-Klassenbenennungen. Damit wurden die Wörter auf ihre unflektierten Stammformen reduziert (sofern verschieden, Bildung von Singular- und Pluralstammformen) sowie mit Wortklasseninformation (z.B. "~~Adj" für Adjektiv, "~~N" für Substantiv) angereichert. Die einzelnen Stammformen wurden mit einer impliziten Trunkierungsvariablen ausgestattet, die es ermöglichen sollte, ein

[4] Vgl.: http://harvest.sourceforge.net/harvest/doc/index.html [07.06.2004]

[5] Über Eigenschaften und Anwendung dieser Version der UDK informiert eine Reihe von Publikationen; zu den neueren davon zählen die Berichte von Schwaninger (1997) und Pika (2002).

[6] Die entsprechenden Angaben dazu variieren in der Projektliteratur.

[7] Dafür wurde Software der finnischen Firma *Lingsoft* (http://www.lingsoft.fi [08.06.2004]) eingesetzt.

Matching mit den konkreten, in den Texten auftretenden Wortformen durchzuführen. Bei kurzen Wörtern musste allerdings darauf verzichtet werden, um das Risiko falscher Übereinstimmungen zu vermeiden (andernfalls würde z.B. die Rechtstrunkierung von "gene" auch Übereinstimmung mit "general", "generic" usw. erbringen).

Schliesslich erfolgte auch noch die Entfernung von Stoppwörtern (Konjunktionen, Präpositionen, Artikel, häufig auftretende Verben und Hilfsverben, Abkürzungen wie "bzw.", "etc.") anhand einer externen, zweisprachigen Stoppwortliste.

Abbildung 5-1 veranschaulicht die Abfolge dieser Schritte anhand eines konkreten Beispiels; dabei werden die deutschen und englischen Klassenbenennungen selektiert, die invertierten Benennung sowie die aufzählenden UND-Verbindungen aufgelöst, Stammformen gebildet und Stoppwörter eliminiert.

Rohdaten aus der UDK (Züricher Version)

```
001Z ~03
002DDUEBERSETZUNGEN / TECHNISCHE U. NATURWISSENSCHAFTLICHE
003DETRANSLATIONS / TECHNICAL AND SCIENTIFIC
004DFTRADUCTION / SCIENTIFIQUE ET TECHNIQUE
```

Ergebnisse des Aufbereitungsprozesses

```
translation~~N/technical~~Adj and~~Conj scientific~~Adj
uebersetzung~~N/technisch~~Adj u.~~Conj naturwissenschaftlich~~Adj
technical translation; scientific translation
technisch uebersetzung; naturwissenschaftlich uebersetzung
```

Abbildung 5-1: *GERHARD* – Aufbereitung der UDK-Einträge (nach Möller et al. 2000)

Abbildung 5-2 zeigt Beispiele für die endgültigen Einträge im UDK-Lexikon – Einzelwörter wie auch Phrasen, die zumeist mit der Zeichenfolge ":-" (trunkiertes Wortende) enden, worauf die jeweils entsprechende Notation folgt. Im Falle des Eintrags "gene" weist die Zeichenfolge ":xxx" darauf hin, dass die Stammform als solche für eine Wortform steht und neben dieser nur die Endung "-s" möglich sein soll.

```
technical translation:-:~03
scientific translation:-:~03
althochdeutsch:-:~30-022
old high german:-:~30-022
althochdeutsch woerterbuch:-:~30-022
althochdeutsch woerterbuecher:-:~30-022
old high german dictionary:-:~30-022
old high german dictionaries:-:~30-022
semantisch information:-:519.767.6
semantisch informationsverarbeitung:-:519.768
gene:xxx s:575.113.1
```

Abbildung 5-2: *GERHARD* – Beispiele für UDK-Lexikoneinträge (nach Möller et al. 1999b; 2000)

Das UDK-Lexikon wurde als "Buchstabenbaum",[8] in einem "Notationserkenner" ("recogniser") implementiert. Neben den Zeichen der Begriffe wurde dabei auch ein Trunkierungssymbol ("#") verwaltet, wodurch es möglich war, bei der Textanalyse spezifische Wortformen auf die reduzierten Lexikoneinträge abzubilden.

5.1.2 Klassifizierungsprozess

Abbildung 5-3 veranschaulicht die Schritte des Klassifizierungsverfahrens. Im ersten Schritt wurden die Textdokumente mit den Einträgen des UDK-Lexikons abgeglichen und auf diese Weise eine Menge von Kandidaten-Notationen ("bag of notations") ermittelt. Im zweiten Schritt wurden aus diesen mit statistischen und heuristischen Verfahren die den Texten schliesslich zugeteilten UDK-Klassen selektiert.

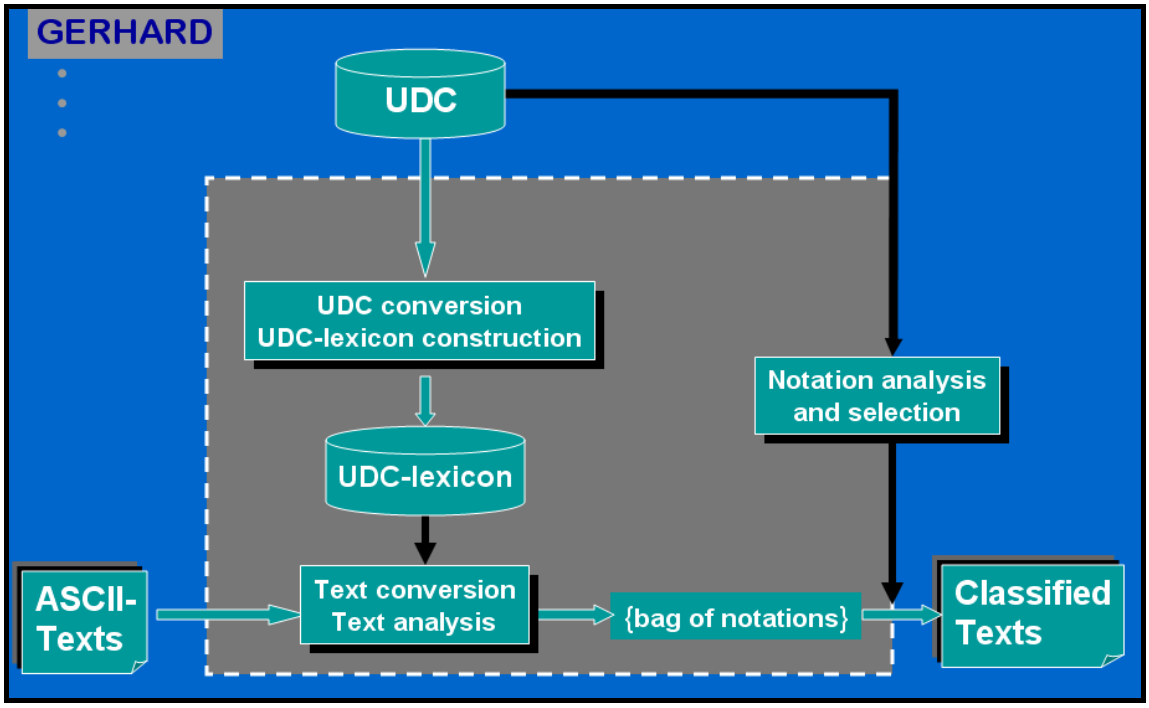

Abbildung 5-3: *GERHARD* – Klassifizierungsprozess
(Quelle: Möller et al. 1999b)

Textaufbereitung. Die zu klassifizierenden Dokumente mussten zunächst an die Zeichenformate des UDK-Lexikons angepasst werden. Dies geschah durch Entfernung der HTML-Kodierung, Auflösung der Umlaute, Umwandlung in Kleinschreibung sowie die Eliminierung von Stoppwörtern (unter Verwendung derselben Stoppwortliste wie im Falle des Lexikons).

Matching. Zur Analyse der Texte wurden diese sequentiell abgearbeitet und iterativ mit den Einträgen des UDK-Lexikons nach dem "longest match"-Prinzip verglichen. We-

[8] Eine in der Computerlinguistik verwendete, effiziente Speicherungsform.

sentliches Charakteristikum dieses Prozesses war die Fähigkeit der Software zur Erkennung von Mehrwortphrasen unter gleichzeitiger Berücksichtigung der Trunkierungsvariablen. So war es etwa möglich, einen Text wie z.B. "Auswirkungen verschiedener Umwelteinflüsse auf Frauen am Arbeitsplatz" auf den Lexikoneintrag "umwelt# frau#:396,5.000.504" abzubilden (Wätjen et al. 1998, 17–18). Als Resultat dieses Analyseprozesses lag pro Dokument eine bestimmte Menge von UDK-Notationen ("bag of notations") mit den zugehörigen textuellen Lexikoneinträgen sowie ihren Auftretensfrequenzen vor.

Statistische Analyse und Gewichtung. Um aus der Menge der durch die Textanalyse einem Dokument zugeordneten Notationen die relevanten zu selektieren, wurde ein Auswahlalgorithmus verwendet, der folgenden Kriterien höhere Bedeutung zuwies:

- Zahl der Notationen mit einem gemeinsamen "Präfix" (je höher diese Zahl, desto sicherer erschien die Zuordnung zum entsprechenden Themenbereich der UDK);
- Länge der Notationen (als Ausdruck ihrer Spezifität);
- Länge der textuellen Zeichenfolge, für die Übereinstimmung gefunden wurde (als Indikator für Spezifität bzw. Eindeutigkeit).

Durch Abschneiden bei einem heuristischen Schwellenwert (durchschnittliche Länge der Text-Übereinstimmungen in Relation zur durchschnittlichen Länge der Notationen) wurde die Zahl der einem Dokument zugeordneten Notationen im Mittel auf 14 gesenkt. In einem weiteren Schritt wurden diese verbliebenen Notationen nach folgenden Kriterien gewichtet:

- Auftretensfrequenz der Notation in Relation zur Länge des Dokuments;
- Höhere Bewertung von Notationen, die auf Basis von Titelwörtern ermittelt wurden, im Vergleich zu solchen, die auf Basis des Textes des Dokuments gewonnen wurden (in einer späteren Phase des Projekts wurden auch noch die HTML-Headings als Unterscheidungskriterium herangezogen);
- Geringere Bewertung geographischer Terme;
- Höhere Bewertung der spezifischeren von zwei aus demselben UDK-Ast vorliegenden Notationen.

Bestimmung der endgültigen Klassen. Durch Anwendung eines (in der Projektliteratur nicht näher erläuterten) heuristischen Schwellenwertes wurde die Zahl der pro Dokument endgültig zugeordneten Klassen schliesslich auf durchschnittlich sechs bis acht reduziert. Die errechneten Gewichte dienten auch für das Ranking der Dokumente bei der Ausgabe (Benutzeroberfläche).

5.1.3 Evaluierung

Im Rahmen eines Testbetriebes nahm das Projektteam Befragungen von Studierenden, Bibliothekaren, Informatikern und anderen Informationsfachleuten vor, was zu einer

Reihe von Verbesserungen führte, wie bspw. der Verringerung der Redundanz in den Ergebnissen durch eine Änderung des Algorithmus, wodurch keine gleichzeitige Zuordnung zu Haupt- und Unterklassen mehr erfolgte (Wätjen et al. 1998, 30). Durch dieses Feedback wurden aber auch einige grundsätzlich problematische Aspekte ersichtlich:

- die Heterogenität der Web-Ressourcen in bezug auf Grösse und Struktur;
- die Dominanz von Naturwissenschaften und Technik in der Züricher Version der UDK;
- die nicht komplette Dreisprachigkeit der Züricher Version der UDK (französischsprachige Klassenbenennungen nur zu 40% vorhanden);
- das aufgrund der Mehrsprachigkeit verstärkte Homonymieproblem (z.B. "windows" als Begriff aus der Datenverarbeitung und aus dem Bauwesen), v.a. bei sehr kurzen Klassenbenennungen.

Der in der Projektliteratur genannte Anteil von 80–85% richtigen automatischen Zuordnungen beruht offensichtlich nur auf Schätzungen des Projektteams.[9]

Ein von Krüger (1999) durchgeführter "Retrievaltest" basierte auf der Inspektion der jeweils ersten 20 erreichbaren Dokumente aus 20 zufällig ausgewählten UDK-Klassen der Gebiete Informatik und Wirtschaftswissenschaften. Diese 400 Dokumente wurden zunächst hinsichtlich ihrer korrekten Zuordnung zur betreffenden Klasse analysiert. Bei korrekter Zuordnung erfolgte auch noch eine Differenzierung nach "relevant" bzw. "nicht relevant", wobei das aus dem *MILOS II*-Retrievaltest bekannte "weit gefasste" Relevanzkriterium zur Anwendung kam (d.h. die Vermutung, ein Benutzer würde sich das Dokument näher ansehen, galt als Indikator für "relevant"; vgl. Sachse et al. 1998, 30–31). Die resultierenden Daten zeigten, dass das Verfahren 83,75% der Dokumente "richtig klassifiziert" hatte[10] – allerdings bei einer erheblichen Varianz zwischen den Beispiel-Klassen, wobei die Fehlklassifikationen primär auf Homonymie zurückzuführen waren. Im Durchschnitt erwiesen sich nur 59% der Web-Ressourcen als "relevant", da die Kollektion zahlreiche Dokumente von institutionsinternem Charakter aufwies. Daneben stellte sich auch heraus, dass fast ein Drittel aller in den Dokumenten enthaltenen Links *veraltet* waren (d.h. nicht mehr funktionierten), was auf ein Aktualitätsproblem der Kollektion hindeutete. Dieser Test zielte also wohl eher auf eine Bewertung des *Dienstes GERHARD* ab und kann m.E. nicht als eine Evaluierung des *Klassifikators* selbst gelten. Weder wurde dabei mit zuvor von Experten klassifizierten Dokumenten gearbeitet, noch wurden die Parameter Recall und Precision in der in Abschnitt 2.6 dargelegten Weise untersucht.[11]

---

[9] In Sekundärpublikationen mutierten solche Schätzungen mitunter auch zu *Fakten*; so z.B. in Voss & Gutenschwager (2001, 305).

[10] Dies galt als erfüllt, "wenn es thematisch zur zugeordneten UDK-Klasse passt[e]" (Krüger 1999, 37); eine Prüfung möglicher besserer Alternativen (anhand der UDK) fand nicht statt.

[11] Die Feststellung Krügers, wonach der Recall nicht zu ermitteln sei, deutet darauf hin, dass die Methodik der Evaluierung eines Klassifikators wohl gar nicht bekannt war.

5.1.4 Benutzung

Ab April 1998 wurde *GERHARD* im Echtbetrieb im WWW angeboten.[12] Das Hauptmenü ermöglichte u.a. das Browsing mittels der UDK, die Suche in den Klassenbenennungen und die Suche im Text der Dokumente. Auf der obersten Stufe wurden die Hauptgebiete der UDK *alphabetisch* angezeigt, auf den weiteren Stufen *systematisch* nach den Notationen. Die Notationen selbst wurden dabei jedoch "bewusst versteckt" und schienen nur in den Links, mit denen in der Hierarchie navigiert wurde, als Suchargumente auf. Für den Einstieg wurde die UDK so modifiziert, dass bestimmte wichtige Gebiete (z.B. Informatik, Wirtschaftswissenschaften) bereits auf der obersten Hierarchiestufe angeboten werden konnten (Wätjen 1998b, 282); auf diese Weise umfasste die oberste Ebene insgesamt 20 Hauptgruppen sowie eine geographische Gruppe. Die Gruppe "Allgemeines", zu der etwa Untergruppen wie "Informationswissenschaft" und "Bibliothekswesen" gehören, wurde allerdings *nicht* auf der obersten Ebene angezeigt, sodass man die genannten Gebiete nur durch eine Recherche in den Klassenbenennungen ansteuern konnte oder bei der Inspektion einer Ergebnisvollanzeige auf einen entsprechenden Link stiess. Wie das Beispiel in *Abbildung 5-4* zeigt, wurden diese Vollanzeigen so gestaltet, dass sie neben Titel, URL usw. auch die dem Dokument zugeordneten UDK-Klassen als Links für das thematische Weiternavigieren aufwiesen. Damit war das Projektziel einer Integration von Suche und Browsing erreicht.[13]

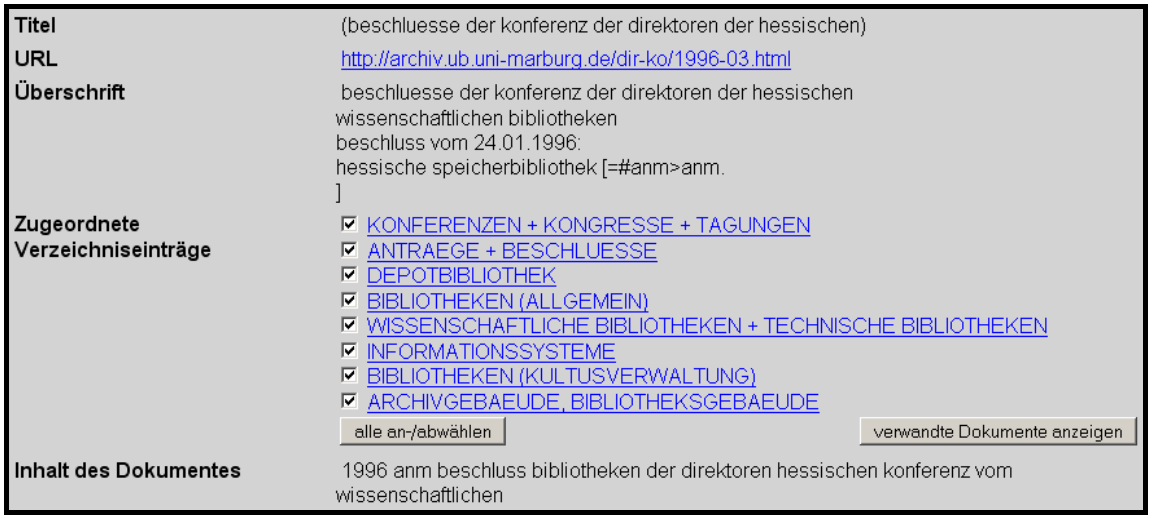

Abbildung 5-4: *GERHARD* – Vollanzeige mit Links zum Weiternavigieren (Quelle: http://www.gerhard.de [13.06.2004])

Zur Benutzungsfrequenz selbst liegen kaum Daten vor. In dem nur zwei Monate nach Eröffnung des Echtbetriebes erschienenen Projektbericht wurden durchschnittlich

[12] http://www.gerhard.de [02.03.2004]

[13] Genauere Beschreibungen der Benutzeroberfläche finden sich in folgenden Publikationen: OFFIS (1998); Wätjen (1998b, 282ff.); Wätjen et al. (1998, 22ff.); Krüger (1999, 15ff.)

2.000 Benutzeranfragen pro Tag, bei steigender Tendenz, angegeben (Wätjen et al. 1998, 31). Aus anderer Quelle geht hervor, dass das Browsing mittels UDK das meistbenutzte Feature von *GERHARD* war; die Zahl der Clicks vor dem Ausstieg betrug im Mittel 11,2 (Möller et al. 1999b, sl. 34). Zu der im Projektbericht angekündigten begleitenden Benutzerforschung und -statistik während des Dauerbetriebes konnte in der Literatur keine weitere Information gefunden werden.

Offensichtlich währte der Dauerbetrieb mit Harvesting, Klassifizierung und Indexierung auch nicht sehr lange. Die gegenwärtig noch abfragbare Webseite von *GERHARD* enthält unter Hinweis auf das Folgeprojekt *GERHARD II* die Information, dass der Datenbestand nicht mehr aktualisiert werde. Die Anzahl der Dokumente wird mit knapp 1,3 Millionen angegeben, die Zahl der UDK-Zuordnungen mit knapp 6,2 Millionen. Die Suche in den Texten der Dokumente wird nicht mehr angeboten.[14]

5.2 GERHARD und DESIRE II

Wie bereits in Abschnitt 3.2.6 erwähnt, wurden auf der Basis einer Kooperation zwischen den Projekten *DESIRE II* und *GERHARD* die Dokumente aus dem skandinavischen Projekt zu Testzwecken mit dem deutschen Verfahren klassifiziert (Koch & Ardö 2000, 21–22). Dabei ging es nicht nur um die Anwendung anderer linguistischer und heuristischer Ansätze, sondern vor allem auch um die Frage, welche Auswirkungen die Verwendung eines universellen Klassifikationssystems im Vergleich zu dem in *DESIRE* verwendeten fachspezifischen Schema erbringen würde. Im Zuge dessen wurden nun 101.082 Dokumente[15] nach dem Verfahren von *GERHARD* klassifiziert und erhielten im Mittel jeweils 11 UDK-Notationen zugeteilt. Durch Abschneiden bei einem Schwellenwert (Gewichtungsfaktor 3 von 10) verblieben durchschnittlich noch 3,9 Notationen pro Dokument.

Interessanterweise verteilten sich die ingenieurwissenschaftlichen Web-Ressourcen aus *AE* über alle neun Klassen der UDK, wenn auch – wie zu erwarten war – zu sehr unterschiedlichen Anteilen. Die Hauptklassen 6 und 5, in denen Technik und Naturwissenschaften hauptsächlich beheimatet sind, erhielten 30% bzw. 23% der vergebenen Notationen. Fast die Hälfte aller Notationen entfiel jedoch auf andere Hauptklassen, darunter 14% auf Klasse 0 (Allgemeines, darunter Informationstechnologie) und 15% auf Klasse 3 (Sozialwissenschaften). Sogar Klasse 9 (Geographie/Geschichte) wies noch 8% der Zuteilungen auf. Die Dokumente waren über 13 Hierarchiestufen verteilt, wobei der Grossteil im Vergleich zum *Ei*-Schema 2 bis 3 Hierarchiestufen tiefer einge-

[14] Stichprobenartige Recherchen des Autors erbrachten ausschliesslich Dokumente mit Datum 1998 oder davor; viele der nachgewiesenen Web-Ressourcen waren überhaupt nicht mehr auffindbar.

[15] Wodurch diese Differenz zu den 132.120 Datensätzen in *DESIRE II* zustande kam, ist nicht bekannt.

ordnet wurde. Die 20 grössten Klassen erwiesen sich als sehr voluminös mit z.T. über 5.000 Dokumenten pro Klasse. Die 200 grössten Klassen umfassten jeweils über 500 Web-Ressourcen, während alle übrigen Dokumente auf 4.000 weitere Klassen verstreut waren. Diese Verteilung wurde als *ungeeignet* für die Verwendung in einem Browsing-System bewertet. Da zahlreiche Dokumente in Klassen ohne offensichtlichen Technik-Bezug eingeordnet worden waren, wurde die Vermutung geäussert, dass dies eine Folge des (im Vergleich zur aufwendigen Vokabularanreicherung in *DESIRE II*) eher limitierten Vokabulars des UDK-Lexikons sei. Für die Klassifizierung einer Kollektion fachspezifischer Dokumente würde das Verfahren von *GERHARD* jedenfalls weiterer Adaptierungen und Verbesserungen bedürfen.

5.3 Das Nachfolgeprojekt GERHARD II

5.3.1 Intentionen

Bereits im Projektbericht zu *GERHARD* waren einige Ergänzungen und Verbesserungen angesprochen worden, wie etwa der Austausch des Harvesting-Roboters (Wätjen et al. 1998, 33–34). Die Gesamtheit der bei der Realisierung von *GERHARD* bzw. während des Echtbetriebes entstandenen Ideen zur Verbesserung des Systems wurde in dem bald darauf erstellten Projektantrag GERHARD II (o.J.) systematisch zusammengestellt. Zu diesen Verbesserungszielen zählten:

Effizienteres Sammeln und qualitative Auswahl der Dokumente

- Ersetzen des Roboters von Harvest, der im Parallelbetrieb zu hohe Systemressourcen verbrauchte, durch den im Rahmen von *DESIRE* entwickelten Harvesters COMBINE;
- Steuerung des Sammelprozesses in Abhängigkeit von der Bedeutung der Web-Domäne und den Aktualisierungs- bzw. Zuwachsraten von Dokumenten und auf Servern;
- Nutzung automatischer Verfahren zur Erkennung von Dokumententypen für das Aussondern irrelevanter Dokumente.

Verbesserung der Klassifizierung

- Dokumenttyperkennung, mit dem Ziel der Anwendung unterschiedlicher Klassifizierungsverfahren auf verschiedene Arten von Dokumenten (Homepages, Linksammlungen, Veranstaltungshinweise, Forschungsberichte, Volltextpublikationen, Graphikseiten);
- Aktualisierung/Vergrösserung des UDK-Lexikons, insbesondere Erweiterung der Klassenbeschreibungen durch Titelstichwörter und RSWK-Ketten aus bibliothekarischen Datenpools (ETH Zürich, GBV, LoC);
- Feinere Analyse der UDK-Notationen durch Nutzung von Verknüpfungszeichen, zeitlichen Schlüsseln und Ortsangaben;

- Verbesserte linguistische Textanalyse durch Portierung und Anpassung eines vom ISIV entwickelten morphologischen Parsers, mit dem Ziel der Verfahrensoptimierung, der Zerlegung unbekannter Komposita (in terminologisch bekannte und mit einer Notation versehene Wörter) sowie der Erkennung zeitlicher und räumlicher Präpositionalphrasen;
- Erkennung der Dokumentensprache und Nutzung dieser Information im Klassifizierungsprozess;
- Verbesserung des Klassifizierungsverfahrens durch Entwicklung neuer Algorithmen für die Notationenanalyse und -selektion.

Verbesserung und Funktionserweiterung der Benutzung

- Entwicklung von Profildiensten (personalisierte Startseiten, Alerting-Dienste, Warenkorb usw.);
- Verbesserung der Ranking-Algorithmen durch Einbeziehung bibliometrischer Informationen (Anzahl der Verweise auf das jeweilige Dokument);
- Schaffung einer RDF-Schnittstelle für den Datentausch mit anderen Nachweisprojekten; vgl. Projekt GERHARD2 (2001).

Qualitätskontrolle

- Bewertung von Klassifizierungsergebnissen durch Fachreferenten.

5.3.2 Gegenwärtiger Entwicklungsstand

Der Projektbeginn von *GERHARD II* verzögerte sich bis zum 1. Januar 2001. Über den gegenwärtigen Status ist nicht sehr viel – und folgendes nur dank persönlicher Informationen (Diekmann 2004; Möller 2004) – bekannt:

- Als Harvester wird nunmehr COMBINE verwendet; damit wurden bisher etwa 1,5 Millionen Dokumente gesammelt (neben HTML- nun auch PDF-Dokumente); geplant ist die Aufnahme von 4,5 Millionen Dokumenten;
- Ein neues Klassifizierungsverfahren, das auf dem Training mit vorklassifizierten Dokumenten beruht,[16] wird eingesetzt – allerdings bisher noch nicht mit zufriedenstellenden Ergebnissen;
- Eine völlig neue Benutzeroberfläche wurde erstellt, die u.a. Skalierbarkeit (z.B. auch für Braille), Kompatibilität mit neueren Web-Standards (XHTML, CSS), Favoritenlisten, personalisierte Startseiten, Alerting-Nachrichten bei Eingang neuer Dokumente zu "abonnierten" Notationen sowie Verzicht auf Graphiken aufweist;
- Die Erkennung der Dokumentensprache und der Nutzung dieser Information im Klassifizierungsverfahren wurde als nicht ausreichend wirksam erkannt und nicht realisiert;
- Die Nutzung bibliometrischer Information für das Ranking der Ergebnislisten wurde aus Kapazitätsgründen zurückgestellt;
- Die RDF-Schnittstelle wird realisiert;
- Mit der Betriebsaufnahme wird noch im Jahr 2004 gerechnet.

[16] Dazu dienen bibliographische Datensätze (Titel, Schlagwörter) aus dem Katalog der ETH Zürich; für das Trainieren des Systems werden 4–6 Wochen benötigt.

Bei dem neuen Klassifikator handelt es sich um *7d Classify*, ein Produkt der Hamburger Softwarefirma *7d*.[17] Dieses Programm führt in der Trainingsphase eine linguistische (Stammformenbildung, Kompositazerlegung, Syntaxanalyse) und statistische Analyse der Beispieldokumente durch, um die Klassenbeschreibungen zu generieren. In der Klassifizierungsphase werden in gleicher Weise Dokumentenbeschreibungen für neu zu klassifizierende Dokumente erzeugt und auf die Klassenbeschreibungen abgebildet. Für jedes Dokument resultiert sodann eine gewichtete Liste von Kandidatenklassen (Sieben-D 2003, 5–7).

[17] http://www.7d-ag.de [14.06.2004]

6 Das Projekt Scorpion von OCLC

6.1 Überblick

Das *Online Computer Library Center* (Dublin, OH, US)[1] ist Nonprofit-Organisation, Katalogisierungskonsortium, vielfältiger Dienstleistungsanbieter auf Bibliotheks- und Informationssektor sowie auch Forschungseinrichtung. Nicht zuletzt aufgrund des Besitzes der Publikations- und Vertriebsrechte an der DDC verfügt OCLC über eine langjährige Forschungs- und Entwicklungstradition auf dem Gebiet der Klassifikation im allgemeinen und hinsichtlich der DDC im besonderen (Mitchell 2001; Mitchell & Vizine-Goetz 2002; Vizine-Goetz 1997b; Vizine-Goetz 2001). Etwa ab der Mitte der 1990er Jahre wurden hiebei auch Fragen im Hinblick auf eine automatische Klassifizierung gestellt (OCLC 2003):

- Können bibliothekarische Klassifikationssysteme wie die DDC und die LCC so adaptiert werden, dass sie für eine automatische Erschliessung insbesondere von Web-Ressourcen und sonstigen elektronischen Dokumenten geeignet sind?
- Kann eine linguistische Bearbeitung der Texte dieser Dokumente dazu beitragen?
- Wie können automatische Klassifizierungsverfahren so verbessert werden, dass sie der intellektuellen Klassifizierung so nahe wie möglich kommen?
- Eignen sich die Ergebnisse für das sachliche Browsing bzw. Suchen, oder auch für das Generieren von Metadaten?
- Sollten Verfahren für das automatische Klassifizieren in manuelle Erschliessungswerkzeuge eingebaut werden?

Vor diesem Hintergrund wurde 1996–2000 das Forschungsprojekt *Scorpion* durchgeführt,[2] das in erster Linie der Gewinnung von Erfahrungen mit dem automatischen Klassifizieren mittels der DDC – später auch der LCC – gewidmet war.

Leider handelt es sich bei *Scorpion* um kein systematisch dokumentiertes Projekt. Zwar liegen zahlreiche Aufsätze, Vorträge und Projektpapiere vor, doch fehlt eine auch nur einigermassen zusammenfassende Gesamtdarstellung.[3] Daher wird im folgenden zunächst ein genereller Überblick über den im Rahmen von *Scorpion* verfolgten Ansatz gegeben und danach versucht, das Klassifizierungsverfahren und die wichtigsten Ergebnisse und Erfahrungen des Projekts zu analysieren.

---

[1] http://www.oclc.org [15.06.2004]

[2] Projekt-Homepage: http://orc.rsch.oclc.org:6109/ [06.02.2004]

[3] Bei der z.B. von Hoffmann (2002, 75) als "grundlegende Arbeit" zitierten Web-Adresse handelt es sich nur um die Projekt-Homepage; die beiden in der Literatur häufig als Projektreferenz angegebenen (und nahezu identen) kurzen Papiere von Shafer (1996; 1997a) beschreiben das eigentliche Projekt in jeweils weniger als fünf Sätzen. Einige Sätze mehr enthält ein kurzer Artikel von Shafer (1997b). Die übrige Projektliteratur beschäftigt sich – teilweise ausführlich – mit speziellen Aspekten wie Dewey-Datenbank, Terminologiefragen, Nachbearbeitung, Evaluierung.

Die Grundüberlegung des im Projekt *Scorpion* verfolgten Ansatzes lautete: Die zu klassifizierenden (Web-)Ressourcen werden als Anfragen ("queries") behandelt, die an eine speziell dafür konstruierte Dewey-Datenbank gerichtet werden, welche als Ergebnis eine ranggeordnete Liste von Fachgebieten (Kandidaten-Notationen) ausgibt.

Als Quelle für die Erstellung dieser Datenbank ("Dewey database", auch "Dewey knowledge base")[4] sollte das bei OCLC Forest Press für die Erstellung der gedruckten und sonstigen DDC-Ausgaben verwendete *Editorial Support System* (ESS) dienen.[5] In diesem verfügt ein Datensatz (= eine Notation) über eine Reihe von Kategorien, die für den Aufbau der Dewey-Datenbank ausgewählt und genutzt wurden. *Abbildung 6-1* veranschaulicht diesen Ansatz.

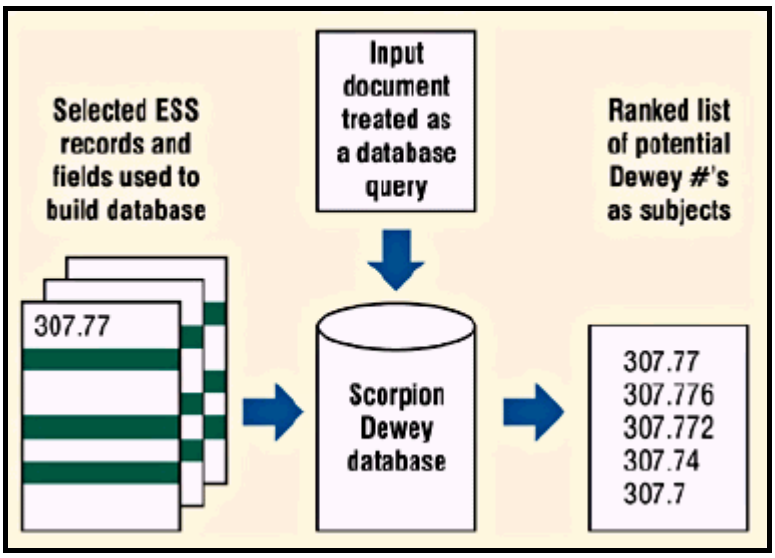

Abbildung 6-1: *Scorpion* – Klassifizierungsprozess
(Quelle: Subramanian & Shafer 1998)

Für die von *Scorpion*[6] ausgegebenen ranggeordneten Kandidaten von DDC-Notationen wurden drei Verwendungsmöglichkeiten ins Auge gefasst:

- automatische Zuteilung, z.B. der ersten *n* Notationen, zum Dokument;
- Eintragung in die Metadaten-Kategorien im Kopf des HTML-Dokuments;
- Präsentation als Vorschläge für einen menschlichen Klassifizierer, der sodann die endgültige Auswahl vornimmt.

[4] Die Übersetzung mit "Wissensbank" wie z.B. bei Koch (1998b, 334) wird hier bewusst vermieden, da dieser Begriff im wissenschaftlichen Kontext primär im Zusammenhang mit Expertensystemen bzw. Decision Support-Systemen verwendet wird.

[5] Version für die 21. Ausgabe der DDC (1996).

[6] Die Bezeichnung *Scorpion* wird im folgenden – wie auch in der gesamten Projektliteratur – sowohl für das Projekt als auch das Verfahren verwendet.

6.2 Die Dewey Datenbank

Wie die in den vorhergehenden Kapiteln behandelten basierte auch das im Rahmen von *Scorpion* angewandte Verfahren auf dem Vergleichen ("matching") von Input-Dokumenten mit dem Vokabular eines Klassifikationssystems. Dieses Vokabular – d.h. die etwa 30.000 mit Notationen versehenen Klassendefinitionen aus dem ESS – musste für die Zwecke des automatischen Klassifizierens aufbereitet und sollte dafür auch weiter optimiert werden. Dies führte zur Untersuchung so verschiedenartiger Fragen wie der folgenden:

- Welche Bestandteile (Felder, Kategorien) der ESS-Datensätze sollen für die Dewey-Datenbank herangezogen werden?
- Wie kann das Vokabular für die einzelnen Klassen weiter angereichert und verbessert werden?
- Wie gut gelingt es der DDC, Information in disjunkte Klassen zu partitionieren; mit anderen Worten, wie verhält es sich mit der Klassenintegrität der Dewey-Begriffsdefinitionen?

Für die verbesserte und angereicherte Version der Dewey-Datenbank, die nicht nur dem Projekt *Scorpion* diente, findet sich in der Literatur auch die Bezeichnung *Enhanced Dewey Database*.

6.2.1 Varianten der Dewey-Datenbank

Bereits im frühen Stadium des Projektes wurde mit 20 bis 30 Varianten[7] der Dewey-Datenbank experimentiert, um deren Effekte im Hinblick auf das automatische Klassifizieren herauszufinden (Shafer et al. 1997). Dabei wurde von einer Grundvariante ("002b") ausgegangen, die nur jene Felder beinhaltete, die für die Erstellung des Produktes *Electronic Dewey for Windows* verwendet wurden. Ohne auf die einzelnen Varianten im Detail einzugehen, kann der folgenden Aufzählung in etwa die Variationsbreite der Massnahmen entnommen werden, die jeweils gesetzt wurden:

- Entfernung von Feldern mit Angaben wie *class elsewhere* oder *see elsewhere*;
- Hinzufügung jeweils aller Klassenbenennungen von der obersten bis zur spezifischsten Hierarchieebene zum Zweck der Disambiguierung problematischer Begriffe (vgl. das Beispiel für "530.1423, Supergravity" in *Tabelle 6-1*);
- Entfernung von Datensätzen mit Notationen, die folgende Attribute aufwiesen:
  - eckige Klammern, z.B. "534[.32]",
  - das Kürzel "vs", z.B. "003.5 vs 629.8",
  - Bereiche wie z.B. "533.1–533.5";
- Entfernung von Feldern, deren Labels mit "m" beginnen ("manual text" Einträge, die menschlichen Klassifizierern bei der unikalen Notationsvergabe helfen sollten und sich beim automatischen Klassifizieren sehr störend ausgewirkt hatten);

[7] Einige davon waren bloss "maintenance refinements" anderer Varianten und daher im Prinzip ähnlich oder gleich wie diese.

- Entfernung von Datensätzen, die das Wort "and" in der Notation enthielten (z.B. "005.269 and 005.284, 005.3684, 005.384");
- Entfernung von Verweisungen zu Einträgen des "Relative Index" der DDC;
- Entfernung der Felder, die Hierarchie-Ergänzungstext aus untergeordneten Ebenen enthielten;
- Ausschalten der Stammformenbildung in verschiedenen Varianten (z.B. überhaupt kein Stemming; kein Stemming ausser Depluralisierung) und Kombinationen mit den oben angeführten Massnahmen.

Dewey Number	Hierarchical Terms
500	Natural sciences and mathematics
530	Physics
530.1	Theories and mathematical physics
530.14	Field and wave theories
530.142	Unified field theory

Tabelle 6-1: *Scorpion* – Hierarchieergänzung zu "530.1423 Supergravity" (nach Shafer et al. 1997)

Welche der Kombinationen bzw. Varianten dieser Massnahmen konkret zu welchen Ergebnissen führte, geht aus der Literatur nur im Zusammenhang mit den weiter unten dargestellten Ergebnissen des Klassenintegritätstests (vgl. 6.2.3) bzw. einiger Evaluierungsergebnisse (vgl. 6.6) hervor. Eine neuere Webseite von OCLC (o.J. a) deutet allerdings darauf hin, dass später das in *Abbildung 6-2* für die Klasse "Robots" (629.892) dargestellte, relativ einfache Format verwendet wurde, das aber immerhin die Benennungen der übergeordneten Klassen sowie zusätzliches Vokabular inkludiert.

```
<rec>
<ScorpionCaption>Robots</ScorpionCaption>
<ScorpionNum>629.892</ScorpionNum>
<ScorpionKeywords>robotics, evolutionary robotics, parallel robots</ScorpionKeywords>
<ScorpionTerms>industrial robots; robotics</ScorpionTerms>
<ScorpionHierarchy>Technology (Applied sciences); Engineering and allied operations;
(Other branches of engineering); Computer control</ScorpionHierarchy>
</rec>
```

Abbildung 6-2: *Scorpion* – Datensatz im Dewey-Datenbankformat (nach OCLC o.J. a, 2)

Dabei haben die als SGML-Tags die Feldinhalte umschliessenden Feldbezeichnungen folgende Bedeutung:

- ***ScorpionCaption:*** Klassenbenennung;
- ***ScorpionNum:*** Notation;

- ***ScorpionKeywords:*** Indexbegriffe, durch DDC-Redakteure zugeordnet;
- ***ScorpionTerms:*** Indexbegriffe, durch statistische Verfahren zugeordnet;
- ***ScorpionHierarchy:*** Vokabular der der jeweiligen Klasse übergeordneten Klassen.

6.2.2 Vokabularanreicherung

Eine bereits erwähnte Form der Vokabularanreicherung, die vor allem der Disambiguierung mehrdeutiger Begriffe dienen sollte, war die Zuordnung der Benennungen der jeweils übergeordneten Klassen. Darüberhinaus wurde mit einer Reihe weiterer Massnahmen versucht, das Vokabular der Dewey-Datenbank zu erweitern bzw. zu verbessern.

Zuordnung von LCSHs. Eine der grundsätzlichen Strategien zur Erweiterung und Aktualisierung der DDC ist "vocabulary mapping", also die Herstellung von Querbezügen zu normiertem Vokabular aus externen Quellen wie den *Library of Congress Subject Headings*, der *Sears List of Subject Headings* u.a. (Mitchell 2001). Seit etwa 1997 wurden diese Aktivitäten unter dem Projektnamen *ExTended Concept Trees* subsumiert (Vizine-Goetz 1997a; 1997b). Insbesondere die LCSH wurden schon seit längerem durch DDC-Redakteure mit Klassen dieses Systems verlinkt (Mitchell 1996); später wurden auch statistische Techniken zur Extraktion solcher Bezüge aus der grössten Datenbank von OCLC, *WorldCat*, entwickelt. Letztere wurden vor allem im Projekt *Popular LCSH with Dewey Numbers*[8] angewandt, in dem die Stärke der DDC-LCSH-Verbindungen in über 700.000 *WorldCat*-Datensätzen mittels eines statistischen Assoziationsmasses analysiert wurde (Vizine-Goetz 1998a; 1998b). Daneben wurde aber auch das Verfahren von *Scorpion* selbst eingesetzt, um LCSH-Datensätze mit Kandidaten-Notationen aus der DDC zu versehen (Vizine-Goetz 1997b).

In diesem Zusammenhang wurden auch Überlegungen angestellt, wie die *Art* der Beziehung zwischen Notation und Subject Heading(s) codiert werden könnte, um die Nutzungsmöglichkeiten in künftigen Werkzeugen für Klassifizierer und Benutzer zu optimieren (Vizine-Goetz 1996). Je nach Zuordnungsmethode wurden folgende Varianten unterschieden (Hickey & Vizine-Goetz 2000, 5):

- IM: Zuordnung durch DDC-Redakteure ("indexer mapped"; stärkste Beziehung);
- SHC: wie IM, doch Zuordnung auf höherer Hierarchieebene (für kürzere DDC-Ausgaben gedacht);[9]
- NF: Zuordnung durch Redakteure von *NetFirst*, einer bibliographischen OCLC-Datenbank von Web-Ressourcen (mittlere Stärke);
- FM: Zuordnung aufgrund der statistischen Auswertung der *WorldCat*-Datenbank ("frequency mapped"; vgl. dazu auch das Feld "Scorpion Terms" in *Abbildung 6-2*).

[8] Dieses hatte zum Ziel, häufig vergebene LCSHs mit DDC-Notationen zu versehen.

[9] SHC = Subject Headings for Children, eine Kurzausgabe der LCSH mit DDC-Notationen.

Durch das intellektuelle Mapping aus den wöchentlichen Ausgaben der LCSH wurden vor allem neuere technische Begriffe wie z.B. "Mixed signal circuits" oder "Nanowires" der DDC zugeordnet. Mitte 2000 waren bereits über 5.400 LCSH durch DDC-Redakteure und weitere 100.000 Sachbegriffe durch andere Techniken zugeordnet worden (Vizine-Goetz et al. 2000, 3).

WordSmith. Im OCLC-Projekt *WordSmith* (1996–2000) wurden linguistisch-basierte Werkzeuge zur Verarbeitung natürlichsprachlicher Texte entwickelt, um daraus Wörter, Phrasen und Begriffe ("concepts") und die Beziehungen zwischen diesen zu identifizieren. Primär ging es dabei um die Extraktion von Mehrwortgruppen ("noun phrases") bzw. Adjektiv-Substantiv-Verbindungen ("noun phrases modified by adjectives"), da viele Begriffe durch solche, im Gegensatz zu Einzelwörtern weniger mehrdeutige Phrasen ausgedrückt werden. Dabei war es besonders bedeutsam, die gebräuchlichen, stabilen Phrasen ("persistent noun phrases") von selten verwendeten oder eher "kreativen" Phrasen zu unterscheiden (Godby & Reighart 1998a; Godby 2001).

Lexikalische Phrase	DDC Notation	Klassenbenennung	Übergeordnete Klasse
gender gap	305.3	Men and women	305 Social groups
Evil Empire	350.5322	Marxism-Leninism	320.53 Collectivism and fascism
Grand Old Party	324.275704	Republican party	324.2 Political parties
Clean Air Act	333.72	Conservation and protection	333.7 Natural resources & energy
corporate welfare	336.243	Corporate income taxes	336.24 Income taxes
jobs programs	336.39	Expenditure	336.3 Public borrowing,debt, expend.
Medicare reform	353.690973	Health insurance	363.6 Admin. of health services

Tabelle 6-2: *Scorpion* – Klassenzuordnungen zu lexikalischen Phrasen (nach Godby & Reighart 1997b)

Phrasen dieser Art, die in bestimmten thematischen Kontexten einen wortähnlichen Status besitzen, wurden "lexical phrases" genannt. Im Zusammenhang mit der Anreicherung der DDC sollten solche lexikalischen Phrasen in grösserem Umfang extrahiert und der DDC zugeordnet werden. In einem von Godby & Reighart (1998b) beschriebenen Versuch geschah dies anhand eines Korpus von aktuellen Nachrichten aus Politik und Tagesgeschehen. Die aus diesen Texten mittels der *WordSmith*-Werkzeuge gewonnenen lexikalischen Phrasen wurden dem *Scorpion*-Verfahren unterzogen und so mit Kandidaten-Notationen aus der DDC versehen. Im Test mit Phrasen, die bereits im Vokabular der DDC enthalten waren, zeigte sich, dass mit der Eingabe der Sätze, in denen die Phrasen vorkamen ("lokaler Kontext"), deutlich bessere Ergebnisse erzielt wurden als bei Verwendung ganzer Textpassagen (Absätze). Auch die (intellektuell vorgenommene) Bewertung der Zuordnung neuerer, aber bereits gebräuchlicher lexikalischer

Phrasen erbrachte sehr zufriedenstellende Ergebnisse. Als Beispiel zeigt *Tabelle 6-2* die von *Scorpion* jeweils an erster Position ausgegebene DDC-Notation zu einigen derartigen Phrasen.

6.2.3 Test der DDC auf Klassenintegrität

Im frühen Stadium des *Scorpion*-Projektes führten Thompson et al. (1997) eine Untersuchung durch, die herausfinden sollte, inwieweit angesichts der Grösse der Dewey-Datenbank das Auftreten überlappender, mehrdeutiger und redundanter Begriffe (Klassen) zu befürchten wäre. Es ging somit um die Fähigkeit der DDC, das zu klassifizierende Material in disjunkte Klassen zu gruppieren, d.h. um die Frage: "What is the *class integrity* of the Dewey concept definitions?" (ibid., 37).

Für diesen Test sollte die DDC ihre eigenen Klassendefinitionen klassifizieren. Dazu wurde jede Klassendefinition – repräsentiert durch den entsprechenden Datensatz aus der Dewey-Datenbank – als "Anfrage" an *Scorpion* gerichtet und dann von diesem Verfahren mit einer ranggeordneten Folge von DDC-Notationen versehen. Dabei kamen sechs der in Abschnitt 6.2.1 erwähnten Datenbankversionen mit jeweils knapp 30.000 Datensätzen zum Einsatz, wobei die Kriterien "Stammformen" (ja/nein) und "Hierarchie" (keine/beide/nur aufwärts) variiert wurden. Als Ergebnismengen wurden jeweils die 20 erstgereihten Notationen betrachtet; unter diesen konnten sich maximal *ein* "self-match" (= die getestete Klasse befand sich unter den 20 erstgereihten des Outputs) und *mindestens* 19 "not self-matches" befinden. Aufgrund des angewandten Gewichtungsschemas konnte allerdings der erzielte Rangplatz nicht isoliert betrachtet werden, sondern es musste auch die Grösse der "tie groups" (d.h. Gruppen von Notationen mit demselben, höchsten Gewicht) mitberücksichtigt werden. Die Resultate waren generell sehr gut, denn

- die überwältigende Mehrheit der Klassen erzielte ein "match" nur mit sich selbst als bestgereihter Klasse (in den Datenbanken mit Hierarchieinformation 97% und höher, in jenen ohne immer noch über 93%);
- in der Datenbank mit Hierarchieinformation (aufwärts *und* abwärts) lagen 99% der Ergebnis-Notationen in "tie groups" von 4 oder weniger Klassen mit dem höchsten Gewicht.

Aus den Resultaten ging des weiteren hervor, dass

- die Stammformenbildung *keinen* wesentlichen,
- das Vorhandensein bzw. Nichtvorhandensein von Hierarchieinformation jedoch einen *starken* Einfluss auf die Ergebnisse beim "self-matching" hatten; dabei war die Variante mit Abwärts- und Aufwärts-Information jener mit nur der letzteren leicht überlegen.[10]

[10] "Downward hierarchy" bedeutete hier die Hinzunahme der Benennungen nur der Klassen auf der jeweils nächstniedrigeren Hierarchiestufe.

Die Analyse der "not self-matches" mit Höchstgewichten zeigte, dass diese bei den Datenbanken mit Hierarchieinformation zu 100% mit der Einerstelle der korrekten Notation übereinstimmten; bei jenen ohne Hierarchieinformation fiel dieses Ergebnis deutlich schlechter aus. Bei den am besten abschneidenden Testdatenbanken stimmten sogar fast 80% *aller* Kandidaten-Notationen (d.h. einschliesslich solcher mit weniger hohen Gewichten) mit der korrekten Notation bis zum Dewey-Dezimalpunkt überein.

Bei der Gesamtinterpretation dieser Testergebnisse wurde festgestellt, dass

- die DDC einen hohen Grad an Klassenintegrität aufweise;
- die Klassen wohldefiniert und disjunkt seien;
- die Anreicherung der Klassen mit hierarchischem Kontext die Fähigkeit von *Scorpion*, Dokumente in einen spezifischeren Notationsbereich einzuordnen, wesentlich erhöhe;
- und dass somit die DDC eine sehr gute Datenbank ("knowledge base") für das automatische Klassifizieren darstelle (Thompson et al. 1997, 43).

6.3 Behandlung der Input-Dokumente

Shafer (1997b) betonte, dass u.a. viel Arbeit für die Vorbereitung der Input-Dokumente aufgewandt worden sei. Gleichwohl finden sich dazu in der Literatur nur zwei wenig detaillierte Informationen.

Zum einen stellte man fest, dass unterschiedliche Arten von Kollektionen auch eine unterschiedliche Vorbehandlung der Dokumente erfordern können. Je nach Dokumententyp könnte es sinnvoll sein, Titel, Headings, Metadaten oder Volltexte bzw. Kombinationen dieser Elemente als "Anfragen" an *Scorpion* zu übergeben. Grössere Dokumente könnten eine Zerlegung in kleinere, spezifischere Teile benötigen – als Beispiel wurden die Artikel einer Enzyklopädie genannt, die sinnvoller einzeln zu klassifizieren wären.

Der zweite Aspekt betrifft das bereits erwähnte *WordSmith*-Instrumentarium, das u.a. zur Extraktion signifikanter Mehrwortphrasen aus natürlichsprachlichen Texten entwickelt wurde. Dieses Verfahren kann nicht nur für die Gewinnung neuer Vokabulareinträge für die DDC genutzt werden, sondern auch dazu, um lexikalische Phrasen aus den zu klassifizierenden Dokumenten herauszulösen. In diesem Zusammenhang entstand die Hypothese, dass Input-Dokumente durch die dabei resultierende Menge solcher Phrasen *besser* repräsentiert würden als durch die Volltexte selbst (Vizine-Goetz et al. 2000, 3). Als Beispiel führte Shafer (1997b; 1997c, sl. 11) einen Versuch an, bei dem mehrere Artikel, die den Terminus "AT&T" aufwiesen, zu *einem* Dokument zusammengefasst wurden, von dem angenommen wurde, dass es eine ausreichende Definition von "AT&T" darstellen würde. Dieses wurde sodann mit *Scorpion* klassifiziert, was jedoch zu einer Zuordnung zu "insurance" führte. Als dasselbe Dokument zunächst mit *Word-*

Smith in Phrasen zerlegt und diese als "Anfrage" an Scorpion übergeben wurden, erfolgte eine Einordnung in "Activities and services" sowie "Telephony".

6.4 Klassifizierungsverfahren

Als Software für die Dewey-Datenbank wurde das bekannte, auf dem Vektorraummodell basierende System *SMART* (Salton & McGill 1987, 125 ff.)[11] verwendet, da dieses für Information-Retrieval-Experimente entwickelt wurde und mit statistisch definierten Dokumentenrepräsentationen und Gewichtungsfaktoren arbeitet. Ausserdem bot es Optionen, mit denen gut spezifiziert werden konnte, welche Teile der ESS-Datensätze jeweils ausgewählt und gewichtet werden sollten (Shafer et al. 1997; Thompson et al. 1997, 38).

Im Rahmen von *Scorpion* wurden verschiedene Gewichtungsschemata ("scoring schemes") aus *SMART* getestet (Shafer & Thompson 1997). Die beiden endgültig verwendeten waren "ATN" und "ATC". Dabei steht:

- ***A*** für "augmented term frequency" (Frequenz, die ein Term aus dem zu klassifizierenden Dokument im Dewey-Datensatz aufweist, in Relation zur Frequenz des häufigsten Terms in diesem Dewey-Datensatz);[12]
- ***T*** für "true inverse document frequency" (Zahl aller Dewey-Datensätze in Relation zur Zahl der Dewey-Datensätze, die einen bestimmten Term aufweisen);[13]
- ***N*** bzw. ***C*** für "no normalization" bzw. "cosine normalization" (diese Normalisierung auf einen Wertebereich von 0,0 bis 1,0 reduziert den "Vorteil", den längere Datensätze im Vergleich zu kürzeren haben).

ATN war somit ein Gewichtungsschema, das die Bedeutung eines Terms in einem bestimmten Dewey-Datensatz, aber auch seine Bedeutung in der gesamten Dewey-Datenbank berücksichtigte. Das Gesamtgewicht ("score"), das einem zu klassifizierenden Dokument pro Klasse zugewiesen wurde, war die Summe der Gewichte für jene Terme, die das Dokument und die betreffende Klasse gemeinsam hatten. ATC war im Prinzip das gleiche, jedoch normalisiert, um den Vorteil längerer Dewey-Datensätze gegenüber kürzen auszugleichen (typischerweise verfügten spezifischere Notationen über mehr Vokabular als allgemeinere).[14]

[11] Vgl. auch: ftp://ftp.cs.cornell.edu/pub/smart [20.06.2004]

[12] Genau: Dieser Wert * 0,5 + 0,5 (dadurch ergibt sich ein Bereich von 0,5 bis 1,0).

[13] Genau: Logarithmus aus diesem Wert.

[14] Die beiden anderen Schemata waren NNN (absolute Termfrequenz im Dewey-Datensatz) und NTC (wie ATC, aber mit absoluter Termfrequenz; Saltons ursprüngliches Gewichtungsschema in *SMART*).

6.5 Nachbearbeitung der Ergebnisse

Obwohl man die Ergebnisse aus dem Klassifizierungsverfahren als generell gut befand, wurden Methoden zur Nachbearbeitung und weiteren Verbesserung der Resultate entwickelt. Diese sollten unerwünschte und unpassende Ergebnisse ("unexpected results", "extraneous results") gleichsam herausfiltern und wurden daher "filters" genannt (Shafer et al. 1999).

Folgende Filtermethoden wurden im Rahmen von *Scorpion* entwickelt bzw. vorgeschlagen:

- ***Three Digits Filter:*** Dieser Filter sollte angewandt werden, wenn eine breitere Klassifizierung angestrebt wird. Die Vorgangsweise besteht ganz einfach darin, alle resultierenden Dewey-Notationen so zu agglomerieren, dass nur dreistellige Notationen resultieren. So würde z.B. aus "341.75" (International economic law) und "341.5" (Disputes and conflicts between states – law) die übergeordnete dreistellige Notation "341" (International law).
- ***Cluster Filter:*** Dabei wurden nicht die Resultate selbst, sondern die Dewey-Datenbank einem Clusterverfahren unterzogen und 495 thematische Cluster, die nicht unbedingt mit Dewey-Hierarchien übereinzustimmen brauchten, extrahiert. Anschliessend an die automatische Klassifizierung einer Dokumentenkollektion mit *Scorpion* sollten die resultierenden Notationen diesen vordefinierten Clustern zugeordnet und jene Cluster, die dabei geringere Gewichte bzw. weniger Einträge erzielen, an das Ende der Ergebnis-Rangreihe gestellt bzw. ausgeschieden werden (Subramanian & Shafer 1998).[15]
- ***Above Average Filter:*** Bei diesem Filter erfolgte die Berechnung des durchschnittlichen Gewichts aller Ergebnisse mit darauffolgender Eliminierung aller Notationen mit Gewichten unter diesem Durchschnittswert.
- ***Additive Caption and Terms Filter:*** Dieser Filter wurde zur Entfernung solcher Notationen entworfen, die nur deshalb ausgegeben wurden, weil der Text des Anfragevektors mit der Beschreibung ihrer übergeordneten Klasse übereinstimmte. Beispielsweise könnte das Verfahren für ein Dokument über "International economic law" neben der Notation für dieses Thema (341.75) auch die spezifischere Notation "341.7582" (International copyright law) erbringen, da die Beschreibung dieser Klasse Wörter enthält, die auch die übergeordnete Klasse beschreiben. In diesem Fall würde die spezifischere Notation weggefiltert werden, da der Anfragevektor kein zusätzliches Kriterium für ihre Vergabe enthielte.[16]
- ***Most General Ancestor Filter:*** Hinter diesem Filter stand die Idee, dass bei Vorliegen mehrerer Notationen aus derselben Subhierarchie und einer übergeordneten Klasse das betreffende Dokument nur dieser übergeordneten Klasse zugeordnet werden sollte. Der Filter subsumiert somit Notationen mit gemeinsamen Präfixen in die jeweils breiteste Klasse unter ihnen. So würde anstelle der Notationen "341.7582" (Copyright), "341.7584" (Design protection), "341.7586" (Patents) und "341.758" nur die letztere Notation (Intangible property) ausgegeben werden.
- ***Collapse Siblings Filter:*** Als Variante des eben beschriebenen Filters würde der gegenständliche auch dann die Notation "341.758" ausgeben, wenn das Resultat *nur* die untergeordneten Klassen enthielte. Hinter diesem Filter stand die Vermutung, dass bei Vorliegen

[15] In dieser Studie war man auch am Vergleich der Leistung von *Scorpion* und dem Clusterverfahren interessiert. Es zeigte sich, dass *Scorpion* bei gleicher Leistungsfähigkeit und Performanz günstiger im Hinblick auf den Verbrauch an Computerressourcen war (ibid.)

[16] Srishaila (2001, 13) hat dies offensichtlich *missverstanden*, da sie schreibt, dass in diesem Fall die breitere Notation weggefiltert würde.

mehrerer gleichrangiger Notationen auch dann das übergeordnete Thema bevorzugt werden sollte, wenn die dafür vorgesehene Notation aufgrund textueller Gegebenheiten im Eingabedokument vom System *nicht* vergeben worden wäre.

Auch eine automatische Vergabe dieser Filter wurde erwogen. Dies hätte jedoch die Entwicklung eines Automatismus bedeutet, der jeweils resultierende Ergebnisvarianten auf ihre Güte bewerten und danach über die Auswahl des/der adäquaten Filtermethode(n) – die zudem auch noch vom Dokumententyp abhängig wäre – entscheiden würde. Diese Idee wurde aber nicht im Detail verfolgt oder gar umgesetzt.

6.6 Evaluierung

Für das *Scorpion*-Projektteam war evident, dass die Güte des automatischen Klassifizierungsverfahrens nur durch einen Vergleich der von diesem ermittelten Notationen ("result set") mit "guten", d.h. denselben Dokumenten *intellektuell* zugeordneten Notationen ("target set") gemessen werden konnte.

6.6.1 Masse für den Vergleich von DDC-Ergebnismengen

Daher wurde eine Liste verschiedener Masse erstellt, mit denen diese beiden Ergebnisse verglichen werden konnten (Shafer 1998; Shafer et al. 1999):

- ***hundreds match*** – Übereinstimmung der ersten Stelle zweier Notationen;
- ***tens match*** – Übereinstimmung der ersten beiden Stellen der Notationen;
- ***ones match*** – Übereinstimmung der ersten drei Stellen der Notationen;
- ***more specific match / more general match*** – wenn eine Notation ein echtes Präfix[17] der anderen ist, so ist die längere Notation ein *more specific match* der kürzeren, während die kürzere ein *more general match* der längeren ist;
- ***correlation match*** – die erste Notation kommt in der Ergebnismenge der zweiten vor, wenn die zweite (d.h. der entsprechende Dewey-Datensatz) mit *Scorpion* klassifiziert wird (vgl. Abschnitt 6.2.3);
- ***synonyms*** – wenn die eine in den Tafeln der DDC als *optionale* Notation für die andere gekennzeichnet ist;
- ***best possible match*** – wenn die intellektuell vergebene Notation das längste Präfix der intellektuell vergebenen Notation in der Dewey-Datenbank von *Scorpion* ist, d.h. wenn *Scorpion* (im Gegensatz zum menschlichen Klassifizierer, der noch weitere Notationsteile anhängen könnte) keine feinere Notation mehr vergeben kann;
- ***exact match*** – völlige Übereinstimmung der beiden Notationen;
- ***close match*** – ODER-Verknüpfung aus *more specific, correlation, synonym, best possible, exact match*;
- ***relevant match*** – wenn irgendeines der obigen Kriterien zutrifft.

[17] Bspw. ist "307.7" ein *echtes* Präfix von "307.77", wohingegen etwa "512.57" und "512.2" nur ein dreistelliges Präfix gemeinsam haben.

Unter Verwendung dieser Masse wurden zwei Ergebnismengen nun so miteinander verglichen, dass eine Notation aus der ersten Menge allen Notationen aus der zweiten Menge gegenübergestellt und danach dieser Prozess für alle anderen Notationen aus der ersten Menge wiederholt wurde.[18]

6.6.2 Die NetFirst-Studie

Mithilfe dieser Masse sollte nun eine Evaluation der Klassifizierungsergebnisse von *Scorpion* vorgenommen werden, wobei als "target set" eine Menge von Datensätzen aus der intellektuell nach der DDC erschlossenen OCLC-Datenbank *NetFirst* herangezogen werden sollte.[19] Jeder *NetFirst*-Datensatz wies mindestens eine und höchstens sieben DDC-Notationen (im Durchschnitt zwei) auf. Eine Auswahl von rund 35.000 der in der Datenbank nachgewiesenen elektronischen Dokumente wurden mit einem Harvester aus dem WWW heruntergeladen und mit *Scorpion* automatisch klassifiziert.

Zu spät wurde allerdings erkannt, dass *NetFirst* für einen solchen Test nicht gut geeignet war, da in dieser Datenbank nur etwa 30% der DDC-Notationen Verwendung fanden. Ausserdem basierte die Notationsvergabe nur auf dem raschen Nachschlagen von Phrasen, die den Inhaltsschwerpunkt charakterisieren sollten, in einem speziell dafür erstellten Register. Eine Überprüfung des Phrasenregisters von *NetFirst* erbrachte, dass diese Phrasen nicht mit der in *Scorpion* verwendeten vollständigen Dewey-Datenbank korrelierten. Daher wurde beschlossen, eine kontrollierte Studie durchzuführen, im Rahmen derer professionelle Klassifizierer eine Kollektion von elektronischen Ressourcen nach der DDC erschliessen sollten. Ob diese Untersuchung tatsächlich durchgeführt wurde und welche Resultate dabei allenfalls zustande kamen, geht aus der Literatur jedoch nicht hervor.

Im Rahmen der *NetFirst*-Studie wurde auch versucht, die Übereinstimmung zwischen "result set" und "target set" zu modellieren. Die erste Graphik ("A") in *Abbildung 6-3* zeigt die schlechte Übereinstimmung zwischen den mit *Scorpion* erzielten Resultaten und der intellektuellen Klassifizierung; diese betrug offensichtlich nur 15% (Godby & Vizine-Goetz 2000, 25). Korrekter sei es jedoch – so wurde argumentiert –, von einem anderen Bild der intellektuellen Erschliessung auszugehen, da bei Klassifizierung derselben Ressource durch eine Reihe von menschlichen Klassifizierern sich deren Urteile nur zu einem gewissen Grad überschneiden würden. Somit wäre das in der mittleren Graphik ("B") dargestellte Bild die Repräsentation der eigentlich zu erwartenden Ergebnismenge aus einer intellektuellen Erschliessung ("expected result set"). Im Vergleich zu den erwähnten 15% exakter Übereinstimmungen betrug der Anteil von "hun-

[18] Ein detailliertes Beispiel für zwei Datensätze findet sich bei Shafer et al. (1999).

[19] Vermutlich hatte OCLC als Anwendung für *Scorpion* ursprünglich die Automatisierung der Erschliessung gerade dieser Datenbank, die ja Web-Ressourcen nachwies, im Sinne.

dreds matches" 90% (ibid.)[20] Das Ziel müsse daher sein, alle Resultate von *Scorpion* innerhalb dieses – wie immer definierten – "expected sets" zu plazieren ("C").

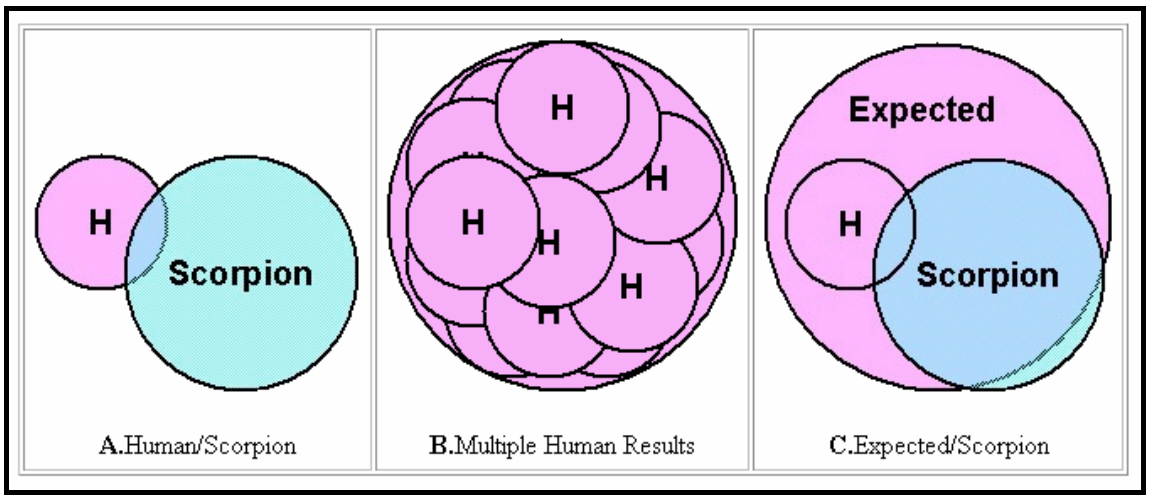

Abb. 6-3: *Scorpion* – Vergleichsdiagramme für Ergebnismengen (Quelle: Shafer et al. 1999)

Die Erstellung eines solchen "expected set" wäre in der Praxis vermutlich relativ schwierig, da die intellektuelle Klassifizierung derselben Kollektion durch zahlreiche verschiedene Personen nicht realistisch ist. Shafer et al. (1999, Kpt. 5) versuchten daher, eine Lösungsmöglichkeit mittels der oben erwähnen Masse zu konzipieren. Dabei wurde von dem Ergebnisset ausgegangen, das *ein* menschlicher Klassifizierer produziert hat; die resultierenden "exact matches" wurden dann um bestimmte andere Übereinstimmungstypen erweitert. Dass dabei die "correlation matches" und die "close matches" einfach dazugenommen werden könnten, mag plausibel erscheinen; die weitere Approximierung durch Einbeziehung von "ones", "tens" und "hundreds" wirkt dagegen ziemlich arbiträr.

Soferne weitere Evaluierungsversuche von *Scorpion*-Resultaten vorgenommen wurden, fanden diese bedauerlicherweise nicht mehr Eingang in Veröffentlichungen.

6.7 Scorpion und DESIRE II

Wie schon in Abschnitt 3.2.6 erwähnt, wurde im Rahmen einer Kooperation zwischen OCLC und NetLab das Verfahren von *Scorpion* – unter Verwendung der manuell und maschinell mit LCSHs angereicherten Dewey-Datenbank – auf die Kollektion aus dem Projekt *DESIRE II* angewandt. Aus der Literatur geht hervor, dass eine ganze Reihe von Testschritten, die auch über die Laufzeit von *DESIRE II* hinausgehen sollten, geplant war (Koch & Ardö 2000b; Koch & Vizine-Goetz 1999; Vizine-Goetz 1999a). Für Net-

[20] Wenn allerdings Übereinstimmung rechts des Dezimalpunktes gefordert war, sank der Anteil auf 60% (ibid).

Lab ging es dabei um die Möglichkeit, Erfahrungen mit einem weiteren System zu gewinnen – die Resultate mit *GERHARD* waren ja wohl nicht sehr zufriedenstellend ausgefallen (vgl. Abschnitt 5.2); OCLC war an einem Test der *WordSmith*-Werkzeuge mit ingenieurwissenschaftlichen Dokumenten und an der Verwendung der DDC bzw. der Dewey-Datenbank für eine fachlich spezialisierte Kollektion mit "deeper-level" Vokabular interessiert.

Interessanterweise stellte sich heraus, dass die DDC sogar über deutlich mehr ingenieurwissenschaftlich relevante Klassen verfügte als das spezialisierte *Ei*-Klassifikationsschema (2.100 vs. 700 Klassen). Andererseits war das in Lund verwendete, mit Thesaurusbegriffen angereicherte Vokabular ungleich reichhaltiger gewesen als jenes der DDC (durchschnittlich 20 Wörter vs. 3 Wörter pro Klasse). Um das grundsätzliche Fehlerpotential abzuschätzen, wurde ein grosses Subset der *EELS*-Dokumente (ca. 50.000) mittels *Scorpion* klassifiziert und pro Dokument eine ranggeordnete Liste von 10 Notationen ausgegeben. Dabei fielen 72% der resultierenden Notation in vier Hauptklassen der DDC: *"000" Computers, Information & Generalities* (27%), *"300" Social Sciences* (28%), *"500" Science* (9%) und *"600" Technology* (28%).[21] Dies wurde als Hinweis darauf gewertet, dass die Zuordnungen mit überwältigender Mehrheit die in den Dokumenten enthaltenen Kontexte reflektierten (Godby & Vizine-Goetz 2000, 24).

Auch bei der Klassifizierung der *EELS*-Dokumente ging man von der Hypothese aus, dass die Genauigkeit der automatischen Zuordnung von Notationen gesteigert werden könne, wenn anstelle des Volltextes der Dokumente nur Surrogate, die die signifikantesten Phrasen enthielten, als Input verwendet würden. Bei der Auswahl der mit *WordSmith* erstellten Phrasen wurden zwei Heuristiken genutzt, die "term stability" bzw. "topicality" messen sollten; jene Phrasen, die Werte über dem Durchschnitt aufwiesen, wurden ausgewählt. Dabei konnte es z.B. vorkommen, dass ein Dokument von 2.300 Wörtern auf nur fünf Phrasen reduziert wurde.

Dewey Class	WordSmith %	Raw Document %
000 Computers, Information & General Reference	31.9	46.3
300 Social Sciences	19.3	16.0
500 Science	15.8	11.5
600 Technology	24.0	19.2
Total	91.1	92.9

Tabelle 6-3: *Scorpion* – DDC-Hauptklassen bei Klassifizierung der *EELS*-Dokumente (nach Vizine-Goetz et al. 2000)

[21] Godby & Vizine-Goetz (2000, 24) führen auch noch Ergebnisse für die zweite Hierarchieebene an.

Beim Vergleich der Klassifizierung auf Basis dieser Phrasen mit den bei der Klassifizierung der Volltexte erzielten Ergebnissen zeigten sich auf der Ebene der einzelnen Dokumente oft nur geringfügige Differenzen. Die Analyse der aggregierten Daten ergab aber doch deutliche Unterschiede (vgl. *Tabelle 6-3*), wobei man – vorbehaltlich einer intellektuellen Überprüfung – von der Annahme ausging, dass die bei Anwendung von *WordSmith* erzielten Klassifizierungsergebnisse die Themen in den echten Dokumenten besser repräsentieren würden (Vizine-Goetz et al. 2000, 3f.)

In beiden Fällen entfielen über 90% der fünf erstgereihten Notationen auf die vier bereits oben erwähnten Hauptklassen; auch das Muster der Verteilung auf die Hierarchieebenen – mit Hauptgewicht auf den Ebenen 3 und 4 – fiel relativ ähnlich aus (ibid.) Auffällig ist allerdings, dass das auf den *WordSmith*-Phrasen basierende Ergebnis eine ausgewogenere Verteilung auf die vier Hauptklassen zeigt als das Resultat der mit den Volltexten vorgenommenen Klassifizierung.

Detailliertere Ergebnisse aus der Anwendung von *Scorpion* auf die EELS-Dokumente wurden nicht veröffentlicht. Auch die anderen von Koch & Ardö (2000b, 22–23) anvisierten Schritte wie die Anreicherung der Dewey-Datenbank mit *Ei*-Vokabular, die Erstellung einer DDC-basierten Browsingstruktur und ein Vergleich dieser mit der *Ei*-basierten wurden offenbar nicht realisiert.

6.8 Scorpion und die LCC

In einer späteren Projektphase von *Scorpion* war OCLC bestrebt, neben der DDC auch die LCC als "knowledge base" für das automatische Klassifizieren zu nutzen (Godby & Stuler 2001). Allerdings sah man diesbezüglich drei Hindernisse:

(a) die kompletten Tafeln beinhalteten fast eine Viertelmillion Klassen;

(b) die Notationen begünstigten die Hospitalität (zulasten von Ökonomie und Konsistenz);

(c) die Tafeln waren nicht für eine automatische Verarbeitung geeignet.

In einer früheren Studie war aber Larson (1992; vgl. Abschnitt 7.1.1) zu der Ansicht gelangt, dass durch eine Vereinfachung der LCC-Tafeln und ihre weitere Anreicherung mit LCSHs die Voraussetzungen für entsprechende Anwendungen gegeben wären. Daher entschloss man sich bei OCLC, für die Zwecke von *Scorpion* die Tafeln der vier für die ersten Tests herangezogenen Teile der LCC – *"Q" Science; "R" Medicine; "S" Agriculture; "T" Technology* – radikal zu vereinfachen, indem 85% aller Klassen mittels einer einfachen Heuristik eliminiert wurden. Dies betraf alle Klassen mit Querverweisungen zu Tafeln anderer Gebiete sowie alle Klassen, die geographische Namen oder Namen von Genres aufwiesen. Dabei nahm man bewusst eine Verflachung der Hierarchie sowie Klassendefinitionen, die durch eine Folge von Notationen (von – bis)

repräsentiert wurden, in Kauf. Die vier genannten Teile umfassten zusammen nur mehr 6.314 Klassen.

Für den Aufbau einer angereicherten LCC-Datenbank wurden LCSHs aus zwei Quellen herangezogen, in denen diese mit LCC-Notationen verknüpft waren – dies allerdings auf spezifischerem Niveau als nunmehr für die vereinfachte Datenbank direkt nutzbar war. Zum einen handelte es sich um das *Library of Congress Subject Authority* (LCSA) File, dessen Notationen durch einen Algorithmus auf die gröbere Hierarchie abgebildet wurden. Zum anderen wurde OCLCs bibliographische Datenbank *WorldCat* genutzt, aus der jedoch die zu einer bestimmten Notation passenden LCSHs nicht so einfach ableitbar waren. Vielmehr war es nötig, ein statistisches Verfahren einzusetzen, das die Stärke der paarweisen Beziehungen zwischen den LCC-Notationen und den LCSHs in den bibliographischen Datensätzen berechnete. Aus diesen Paarbeziehungen wurden die höchsten 9% verwendet, d.h. die LCSHs den entsprechenden LCC-Klassen zugeordnet. In Summe entstammte aber dennoch der grösste Teil der Zuordnungen aus der *WorldCat*-Datenbank und nur ein kleiner Teil aus LCSA.

Mit der resultierenden LCC-Datenbank wurde ein ähnlicher Test durchgeführt wie zuvor schon mit der Dewey-Datenbank (vgl. Abschnitt 6.2.3), indem die einzelnen Datensätze als Anfragevektoren für eine automatische Klassifizierung mit *Scorpion* herangezogen wurden. Für diesen Test wurden drei Datenbanken verglichen:

(a) ***Grunddatenbank***; enthielt nur Vokabular aus der LCC und aus LCSA;

(b) ***Testdatenbank 1***; enthielt zusätzlich zu (a) alle in *WorldCat* gefundenen LCSHs;

(c) ***Testdatenbank 2***; enthielt zusätzlich zu (a) nur die LCSHs aus *WorldCat* mit hohen Assoziationskoeffizienten.

Die Auswertung der Klassifizierungsergebnisse zeigte, dass die Testdatenbank 2 beim "self-matching" mit Abstand das beste Ergebnis erzielte; in 91,1% der Fälle wurde die richtige Klasse an die erste Stelle gereiht. Die Grunddatenbank fiel dagegen deutlich ab (68,5%); die erste Testdatenbank war erwartungsgemäss durch hohen Ballastanteil gekennzeichnet (27,3%). Dieses Resultat wurde als Indikator dafür gesehen, dass LCSH aus einer externen Quelle gut geeignet wären, zur Präzisierung der Klassendefinitionen beizutragen.

Über diesen Pilottest hinaus wurden keine weiteren Erfahrungen mit der LCC-Version von *Scorpion* mehr publiziert. Nach OCLC (o.J. a) kam schliesslich das in Abschnitt 6.2.1 gezeigte einfache Datenformat von *Scorpion* auch für die LCC-Datenbank zur Anwendung.

6.9 Benutzungsmöglichkeiten

Im Gegensatz zu den Projekten aus Lund, Wolverhampton und Oldenburg wurde von OCLC keine grössere, mit *Scorpion* automatisch klassifizierte Dokumentenkollektion mit einer klassifikationsbasierten Browsingstruktur via WWW öffentlich zugänglich gemacht. Dennoch gibt es drei Möglichkeiten der Benutzung von *Scorpion*, die im folgenden kurz skizziert werden sollen.

6.9.1 CORC / Connexion

CORC (Cooperative Online Resource Catalog) war ein 1998 gestartetes und 2000 auch in den Produktionsbetrieb von OCLC integriertes Projekt, im Rahmen dessen ein Verbundkatalog digitaler Ressourcen in weltweiter Kooperation aufgebaut und via WWW zugänglich gemacht werden sollte. Als Werkzeuge für die sachliche Erschliessung kamen *WordSmith* (für freie Schlagwörter) und *Scorpion* (für DDC-Notationen) zum Einsatz; beide sollten die Katalogisierungsarbeit durch Vorschläge unterstützen, d.h. nicht vollautomatisch arbeiten. Die Nutzung von *Scorpion* war dabei an das Abonnement der DDC-Online-Version *WebDewey* gebunden. Anfang 2001 betrug die Zahl der mitarbeitenden Bibliotheken fast 500 (OCLC 2001); im Juni 2002 umfasste CORC 673.259 Katalogisate (Cremer & Neuroth 2002, 266).

Zur Qualität von *Scorpion* im Rahmen von CORC bemerkten Hickey & Vizine-Goetz (2000, 3): "Although the current system often suggests several class numbers which are either exactly or very close to the proper numbers, human review of the numbers is still needed." Die menschliche Intervention wurde vom Katalogisierungssystem in der Form unterstützt, dass die vorgeschlagenen Notationen bzw. Klassenbenennungen Links aufwiesen, die eine nähere Inspektion des entsprechenden DDC-Datensatzes ermöglichten (ibid.)

Seit 2002 ist CORC ein Teil des neuen, integrierten Katalogisierungsdienstes *Connexion*, "OCLC's flagship cataloging service" (OCLC o.J. b). Für *WebDewey*-Abonnenten mit OCLC-Katalogisierungsberechtigung besteht dort nach wie vor die Möglichkeit, bei der Katalogisierung elektronischer Ressourcen mittels *Scorpion* erstellte Kandidaten-Notationen abzurufen bzw. diese dem Katalogisat zuzuweisen (OCLC o.J. c, 1).

6.9.2 WWW-Klassifikatoren

Auf der nach wie vor bestehenden Webseite des *Scorpion*-Projektes[22] werden u.a. Links zu interaktiven Klassifikatoren angeboten, von denen jedoch nur jener für die LCC ohne Registrierung öffentlich zugänglich ist.[23] Das Input-Formular für diesen Dienst (vgl.

[22] http://orc.rsch.oclc.org:6109/ [06.02.2004]

[23] http://orc.rsch.oclc.org:6109/classify/pub/index.html [20.06.2004]

Abbildung 6-4) erlaubt die Auswahl von vier der Hauptgruppen der LCC und die wahlweise Eingabe der Web-Adresse oder des Textes des klassifizierenden Dokuments.

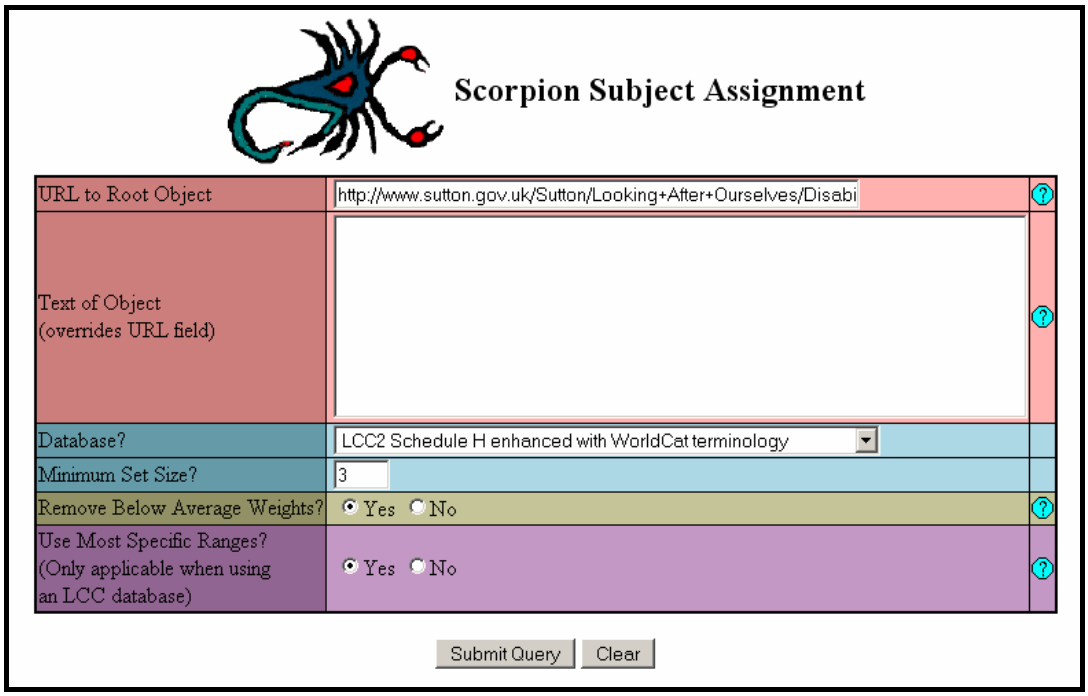

Abbildung 6-4: *Scorpion* – WWW-Klassifikator (LCC), Eingabeformular

Zahlreiche Versuche des Autors mit diesem WWW-Klassifikator erbrachten zum überwiegenden Teil wenig zufriedenstellende Ergebnisse. Ein Beispiel für ein akzeptables Resultat ist in *Abbildung 6-5* (zu klassifizierendes Web-Dokument) und *Abbildung 6-6* (Ergebnis von *Scorpion*) dargestellt.

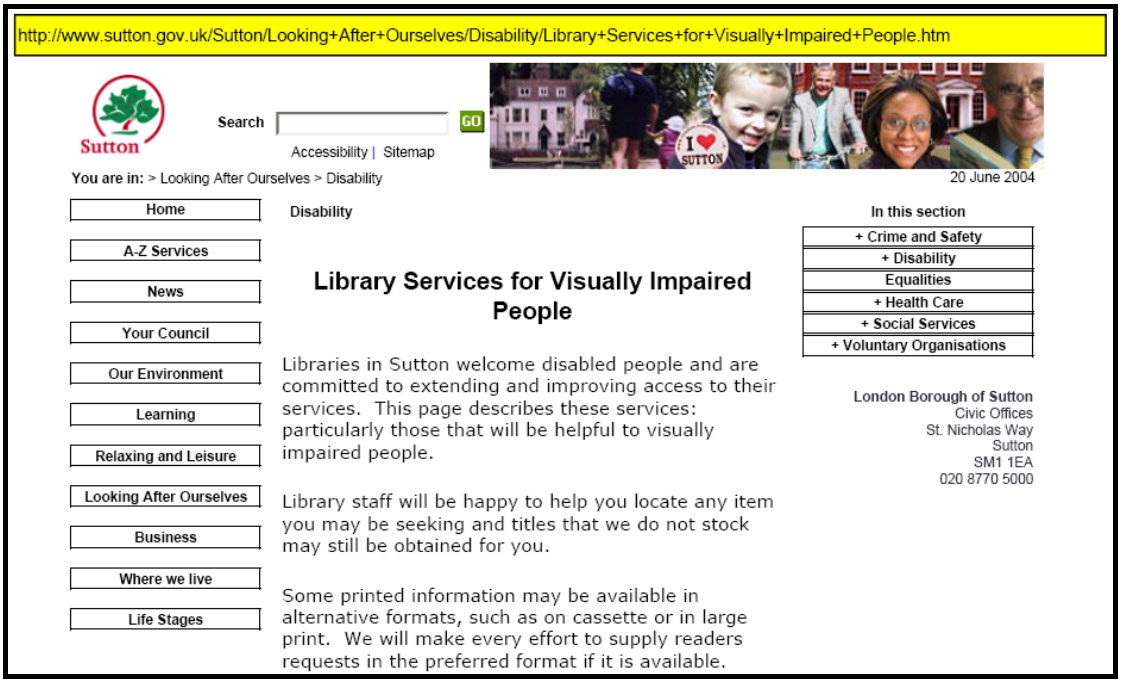

Abbildung 6-5: *Scorpion* – WWW-Klassifikator (LCC), Inputdokument

Weight	Subject Code Range	Subject Text
144	HV1721 -- HV1756	Social pathology. Social and public welfare. Criminology ; Protection, assistance and relief ; Special classes ; Handicapped ; Blind ; Books for the blind. Talking books
108	HV1666 -- HV1698	Social pathology. Social and public welfare. Criminology ; Protection, assistance and relief ; Special classes ; Handicapped ; Blind ; Education of the blind ; Practice ; Systems of typography, writing, etc.
104	HB3718 -- HB3728.52	Economic theory. Demography ; Business cycles. Economic fluctuations ; Relation to special topics

Abbildung 6-6: *Scorpion* – Ergebnis des WWW-Klassifikators (LCC)

6.9.3 Open Source Version

Ebenfalls von der *Scorpion*-Homepage führt ein Link zu einer Download-Seite,[24] von der eine *Open Source*-Version der Software, die auf UNIX-Rechnern lauffähig ist, heruntergeladen werden kann. Diese Version ist für die Verwendung durch Forscher gedacht, die über ein maschinenlesbares Klassifikationssystem verfügen und dieses in ein automatisches Klassifizierungsverfahren einbinden wollen.

[24] http://www.oclc.org/research/software/scorpion/ [10.03.2004]

7 Weitere Anwendungen und Projekte

7.1 Automatisches Klassifizieren von Büchern

Alle in den vorhergehenden Kapiteln betrachteten Projekte widmeten sich dem automatischen Klassifizieren von elektronischen bzw. Web-Dokumenten. Mit der automatischen Zuteilung von Notationen zu *Büchern* bzw. Bibliotheksbeständen haben sich bislang keine grossen Projekte beschäftigt. Die Recherchen erbrachten ausser der vielzitierten und im folgenden referierten Studie von Larson (1992) aber auch einige kleinere Arbeiten zur Automatisierung des bibliothekarischen Klassifizierungsprozesses, die danach betrachtet werden sollen.

7.1.1 Die LCC-Studie von Larson

Die 1992 publizierte Untersuchung von Larson, in der über Experimente zur automatischen Klassifizierung von MARC-Datensätzen nach der LCC berichtet wird, gilt als "landmark study" (Godby & Stuler 2001, 1), da sich zuvor – und auch danach! – niemand in so systematischer Weise mit der automatischen Zuteilung von Notationen eines bedeutenden Klassifikationssystems zu *Büchern* (repräsentiert durch bibliographische Datensätze) beschäftigt hat.

Den Experimenten lagen rund 30.000 MARC-Katalogisate aus der bibliothekswissenschaftlichen Fachbibliothek der University of California (Berkeley), zugrunde, die aufgrund dieser Herkunft zu 92% eine (Aufstellungs-)Notation aus der LCC-Hauptklasse "Z" (Bibliography, Library Science and Information Science) aufwiesen. Diese Datensätze verteilten sich auf 5.765 verschiedene Klassen aus "Z", für die Klassendefinitionen in Form von "classification clusters" erstellt wurden, d.h. Vektoren von Attributgewichten auf der Basis aller *Dokumente*, die zur jeweiligen Klasse gehörten; die Terme aus den Tafeln bzw. Registern der LCC wurden *nicht* verwendet. Die Klassifikationsexperimente wurden mit einer Testmenge von 283 neuen, bereits nach der LCC klassifizierten, aber noch nicht in der Datenbank enthaltenen Katalogisaten durchgeführt. Für das automatische Klassifizieren dieser Dokumente wurde das probabilistische Retrievalsystem *Chesire* eingesetzt (d.h., der Klassifizierungsprozess wurde als IR-Prozess definiert). Ziel war es, für die neuen, ebenfalls durch Vektoren von Termgewichten repräsentierten Dokumente die jeweils beste Klasse ("classification cluster") zu finden bzw. die Übereinstimmung dieses Ergebnisses mit der "wahren" (vorab intellektuell zugeteilten Klasse) zu prüfen.

Zur Bestimmung der Ähnlichkeitswerte für die Rangreihung der Klassen nach dem Grad ihrer Übereinstimmung mit einem Eingabedokument wurde das Skalarpro-

dukt von Klassenvektor und Dokumentenvektor errechnet. Dabei wurden vier verschiedene Methoden der Berechnung von ***Termgewichten*** getestet:

(1) *Coordination level matching:* Termgewichte sind "1" (für das Auftreten eines Terms) und "0" (Nichtauftreten); die endgültige Masszahl ist nichts anderes als die Zahl der im Klassenvektor und im Dokumentenvektor gemeinsam auftretenden Terme.

(2) *TFIDF-Gewichtung:* Diese bekannte Methode gibt den Termen, die in einem Dokument oft, in der gesamten Kollektion jedoch selten auftreten, die höchsten Gewichte, und denjenigen Termen, die im Dokument selten, in der Kollektion jedoch häufig auftreten, die niedrigsten Gewichte.

(3) *"Model 1C":* Dieses probabilistische Gewichtungsverfahren basiert auf der bedingten Wahrscheinlichkeit, mit der ein Klassifizierer für ein Dokument, das den Term T enthält, die Klasse K zuteilen würde. Dies wird auf der Basis der vorab klassifizierten Dokumente für jeden Term berechnet (Klassenvektor). Der Dokumentenvektor besteht aus binären Gewichten wie bei (1).

(4) *Weighted relative frequency matching:* Wie (1), doch erweitert um die relative Häufigkeit des betreffenden Term im Vokabular der Klassendefinition.

Des weiteren wurden fünf Varianten der ***Attributauswahl*** getestet:

(a) *All elements:* Die Terme aus dem Titel und aus allen LCSHs;

(b) *Title and first subject:* Titel und erster LCSH (die Regeln der LoC besagen, dass die LCC-Notation auf der Grundlage des ersten LCSH vergeben werden soll; dies wird allerdings in der Praxis nicht immer befolgt);

(c) *All subjects:* Alle LCSH, aber nicht die Terme aus dem Titel;

(d) *First subject only:* Nur die Terme aus dem ersten LCSH-Eintrag;

(e) *Title only:* Nur die Terme aus dem Titel.

Schliesslich beinhaltete die Versuchsanordnung auch zwei ***Stemming***-Methoden sowie einen Ansatz zur Normalisierung von ***Phrasen***:

(i) *Full stemming:* Die Stichwörter aus dem Titel- bzw. LCSH-Feld wurden auf ihre Stammformen reduziert, wobei das Verfahren aus dem System *SMART* eingesetzt wurde (vgl. Salton & McGill 1987, 139);

(ii) *Plural stemming*: Mit einem einfachen Verfahren wurden die im Englischen gebräuchlichen Pluralformen auf die jeweilige Singularform zurückgeführt (nicht jedoch schwierigere Fälle wie z.B. "thesauri").

(iii) *LCSH phrases*: Die einzelnen LCSH wurden in Kleinbuchstaben umgesetzt und durch Entfernung aller Punkte, Kommas und Leerzeichen in lange Zeichenfolgen verwandelt, die als *ein* Term behandelt wurden; in den obigen Fällen (a) und (b) wurden die Terme aus dem Titel mit "plural stemming" behandelt.

Für die Tests wurden aus den Kombinationen der Termgewichtungsverfahren (1) bis (4) und der Selektionsverfahren (i) bis (iii) insgesamt 12 getrennte Datenbanken von Klassenvektoren erstellt. Die 283 Testdokumente wurden mittels der Kombination der Selektionsverfahren (a) bis (e) und (i) bis (iii) in 15 Anfrage-Mengen transformiert. Dadurch ergaben sich 4 * 5 * 3 Klassifizierungsverfahren bzw. -tests.

Aus der detaillierten Ergebnisdarstellung bei Larson (1992, 138–147) können hier nur die Hauptresultate angeführt werden:

Die besten Ergebnisse – definiert als grösster Anteil der "wahren" Klasse auf Rangplatz 1 – erwies sich die Kombination von *weighted relative frequency matching* (4), *first subject heading* (d) und *plural stemming* (ii). Damit konnten 46,6% der neuen Katalogisate korrekt zugeordnet werden. Eine – allerdings nicht systematisch durchgeführte – Inspektion der Klassen "zweiter Wahl" ergab, dass viele davon eine akzeptable Alternative zu der durch die menschlichen Klassifizierer gewählten Klassen waren. 74,4% der "wahren" Klassen befanden sich unter den jeweils 10 bestgereihten.

In ca. 45% der mit diesem Verfahren korrekt klassifizierten Fälle enthielten auch die (Tafeln der) LCSH die exakte Spezifizierung der betreffenden Notation, während in den übrigen Fällen nur unvollständige Notationen ("partial class numbers"), Bereiche von Notationen (von-bis-Folgen) oder gar keine Notationen verzeichnet waren.

Der Versuch, die LCSH in Form von *Phrasen* zu verwenden (iii), erbrachte keine guten Ergebnisse, da dieser Ansatz zwar viele nicht korrekte, aber auch viele korrekte Klassen zurückwies.

Die auf den *TFIDF-Gewichten* basierende Matching-Methode (2) erbrachte auffallend *schlechte* Resultate. Larson (1992, 146) führte diesen Befund auf die Begrenztheit des Vokabulars zurück, zumal die verwendeten MARC-Datensätze nur Titel und LCSHs enthielten. Im Gegensatz dazu hatte die traditionelle IR-Forschung, in der sich der TFIDF-Ansatz bewährt hatte, meist auf Datensätze zurückgegriffen, die neben dem Titel auch Abstracts und oft eine grosse Zahl von Deskriptoren aufwiesen.

Die verschiedenen Verfahren erbrachten sehr unterschiedliche Resultate in bezug auf die jeweils korrekt klassifizierten Fälle. Eine Analyse dieser "best ranks" ergab, dass 76,3% korrekte Zuordnungen erreichbar gewesen wären, wenn für jedes Buch das jeweils dafür am besten geeignete Verfahren wählbar gewesen wäre.

Larson (ibid.) schloss daraus, dass eine vollautomatische Vergabe von LCC-Notationen nicht realistisch sei, dass aber eine semi-automatische Vorgangsweise, bei der – möglicherweise auch auf der Basis der Kombination einiger Verfahren – ein menschlicher Klassifizierer Vorschläge zur Auswahl dargeboten bekommt, weiter verfolgt werden sollte.

7.1.2 Das ACS-Verfahren von Cheng & Wu

Cheng & Wu (1995), Hongkong Polytechnic University, berichten von einer Untersuchung mit der auch in Südostasien weitverbreiteten DDC. Das von ihnen erstellte *Automatic Classification System* (ACS) basiert auf dem Vektorraummodell und einem von den Autoren neu entwickelten Ähnlichkeitskoeffizienten. Für die Erstellung der Klassenrepräsentationen wurde nicht das Vokabular der DDC herangezogen, sondern jenes der Titel und Kapitelüberschriften einer Kollektion zuvor intellektuell klassifizierter Bücher. Dabei erfolgte die Entfernung von Stoppwörtern und Wörtern mit geringer (d.h.

nur einmaliger) Auftretensfrequenz sowie die Bereinigung von Synonymformen (Akronyme, Schreibvarianten, Singular-/Pluralformen), jedoch keine Stammformenbildung.

Der Algorithmus von *ACS* basierte auf dem erwähnten Ähnlichkeitsmass, wobei die Übereinstimmung eines zu klassifizierenden Buches mit dem Eigenschaftsvektor der jeweiligen Klasse geprüft wurde. Bei über- bzw. untergeordneten Klassen wurde paarweise untersucht, mit welcher Hierarchieebene eine stärkere Assoziation vorlag; war dies die übergeordnete Klasse, so brach das Verfahren ab, war es die untergeordnete, so erfolgte eine weitere Iteration mit der nächsten Hierarchiestufe.

Tabelle 7-1 zeigt die der (eher bescheiden konzipierten) Untersuchung zugrundegelegte Versuchsanordnung sowie die Hauptergebnisse. Bei den untersuchten Dewey-Klassen handelte es sich um "510" (Mathematics) und, als untergeordnete Klasse, "515" (Calculus and analysis). Für beide Klassen wurden aus den Beständen einer Hochschulbibliothek in Hongkong bereits nach der DDC[1] klassifizierte Werke ausgewählt und in zwei Gruppen geteilt. Anhand der Bücher der ersten, grösseren Gruppe ("experimental group") wurden die Klassenrepräsentationen erstellt, während es sich bei den Büchern der zweiten Gruppe ("validation group") um Dokumente handelte, deren Vokabular nicht in diese Klassendefinitionen eingegangen war und die von *ACS* neu zu klassifizieren waren.

	Klasse	Bücher	Korrekt
Gruppe 1	510	238	82,8%
	515	146	91,8%
	Total	384	86.7%
Gruppe 2	510	20	85,0%
	515	20	95,0%
	Total	40	90,0%

Tabelle 7-1: *ACS* – Versuchsanordnung und Resultate (nach Cheng & Wu 1995, 295)

Die Resultate fielen für beide Gruppen mit 86,7% bzw. 90% korrekten Zuordnungen sehr vorteilhaft aus. Um das beste Ergebnis zu erzielen, war auch mit einer Heuristik experimentiert worden, die ein zu klassifizierendes Buch dann der "parent class" zuordnete, wenn der Assoziationskoeffizient für diese Klasse, multipliziert mit einem Schwellenwert, grösser als der entsprechende Koeffizient für die untergeordnete Klasse war. Dabei wurden mit dem besten Schwellenwert (1,20) durchschnittlich 88,4% korrekte Zuordnungen erreicht. Die Fehlklassifikationen wurden auf vier mögliche Faktoren (Fehlen von Kapitelüberschriften, nicht aussagekräftige Titel, in generelleren Wer-

[1] 20. Ausgabe (1987).

ken versteckte Spezialthemen, Fehler der intellektuellen Klassifizierung) zurückgeführt. Für die praktische Anwendung wurde die Kombination mit menschlicher Intervention – Auswahl der Hauptklasse, bei der *ACS* beginnen sollte, Eingreifen bei problematischen Entscheidungen zwischen Klassen – empfohlen.

Obwohl die bei dieser Studie verwendete Kollektion klein und die Zahl der einbezogenen Klassen minimal war, weisen die Resultate darauf hin, dass mit *ACS* ein zwar sehr einfach anmutender, aber durchaus nicht uninteressanter Ansatz vorgelegt wurde.

7.1.3 Sonstige Anwendungen und Projekte

AutoBC. "Automatic Book Classification", ein von Kim & Lee (2002) in Südkorea entwickeltes Verfahren, verwendet die auf S. R. Ranganathan zurückgehende *Colon Classification* (CC), eine in Europa – zumindest im Detail – wenig bekannte Facettenklassifikation. Realisiert wurde vorerst nur ein Klassifikator für das Fachgebiet "Bibliothekswissenschaft". Grundmodul des Systems ist eine nach Fachgebieten und Facetten strukturierte Vokabulardatenbank ("knowledge base"); für das genannte Fachgebiet umfasst diese 387 aus der der CC[2] und der DDC[3] extrahierte Begriffe. Kim & Lee hängen der m.E. eher fragwürdigen These an, wonach die Titel von Monographien alleine genügend Aussagekraft für eine automatische Klassifizierung besitzen würden; nur in Einzelfällen müssten von bibliothekarischer Seite zusätzlich Schlagwörter hinzugefügt werden. *AutoBC* ordnet das (Titel-)Vokabular der zu klassifizierenden Bücher aufgrund der in der "knowledge base" vorgefundenen Strukturen und der Termfrequenzen den Fachgebieten, Facetten und Isolaten der CC zu und kombiniert diese nach der Facettenformel der CC zu gültigen Notationen. Ein Test mit 365 Büchern ergab, dass 81% davon "klassifizierbar" waren und die restlichen nach einer weiteren Anreicherung der Vokabulardatenbank klassifiziert werden konnten. Details über die Güte dieser Zuordnungen sind ebensowenig bekannt wie Einzelheiten über die tatsächliche Vorgangsweise im Rahmen dieses m.E. völlig unzureichend dokumentierten Verfahrens.

ACN und UDC-AUTCS. Diese beiden in Japan erstellten Verfahren sollen Klassifizierer in Bibliotheken bei der Bildung von Notationen maschinell unterstützen. Das bereits in den 1980er Jahren entstandene Modul "Automatic Classification Numbering" (Ishikawa 1988) verwendete die *Nippon Decimal Classification* (NDC) und beruhte auf der interaktiven Eingabe von Sachbegriffen durch den Katalogisierer. Das System suchte diese Begriffe in einer Datenbank, die das Vokabular der NDC (Tafeln, Hilfstafeln und

[2] 7. Auflage (1987)
[3] 21. Ausgabe (1996)

Register) beinhaltete, und schlug Kandidaten-Notationen vor. Nach einer menschlichen Auswahlentscheidung bildete es unter Beachtung der NDC-spezifischen Verknüpfungsregeln die endgültige Notation in formal korrekter Form. Ein "Test" mit 24 [!] bereits manuell klassifizierten Büchern erbrachte eine hohe Übereinstimmungsrate; die übrigen Fälle wurden auf Variation im menschlichen Entscheidungsverhalten zurückgeführt.

Das später für die UDK entwickelte "UDC Number Automatic Combination System" arbeitete nach demselben Schema (Ishikawa et al. 1994). Offensichtlich vermochte diese neuere Variante aufgrund von Eingabeparametern zu erkennen, ob der betreffende Begriff in den Haupt- oder in den Hilfstafeln nachgeschlagen werden sollte. Die endgültige Notation wurde nach einer neuerlichen Bearbeiterentscheidung aufgrund der Verknüpfungsregeln der UDK erstellt.

Woodward (1996, 197) meinte, dass *UDC-AUTCS* wegen Übersetzungsproblemen schwer zu beurteilen sei.[4] Nach m.E. wird jedoch – auch aufgrund der älteren Publikation – klar, dass der von Ishikawa et al. (1994) verfolgte Ansatz wenig mit automatischer Klassifizierung im engeren Sinn zu tun hat und im günstigsten Falls als Katalogisierungshilfe zu bewerten ist.

NDC und Bücher. Ishida (1998) berichtet über ein Experiment, im Rahmen dessen japanische Bücher auf der Basis von Katalogisierungsdatensätzen automatisch den Klassen der NDC zugeordnet wurden. Dabei wurden verschiedene Extraktions- und Gewichtungsmethoden getestet. Der Studie lag eine Kollektion von 1.000 Büchern zugrunde; das beste erzielte Resultat lag bei 55,9% korrekten Zuordnungen.[5]

Automatic Dewey Decimal Classification. Unter dieser Projektbezeichnung versucht eines der bekanntesten LIS-Institute Indiens (Documentation Research & Training Centre, Bangalore), Bücher automatisch nach der DDC zu klassifizieren (Srishaila 2001, 7–8). Das System basiert auf einem Parser für natürliche Sprache, einem Expertensystem und einem regelbasierten Algorithmus für die Notationsvergabe. Das der syntaktischen Analyse der Buchtitel [!] durch den Parser zugrundeliegende Lexikon wurde auf der Grundlage des "Relative Index" der DDC erstellt.[6]

Notationsübernahme und -zusammenführung in OPACs. Der Vollständigkeit halber seien noch einige in der Bibliothekspraxis angewandte Techniken angeführt, die zwar nicht zum automatischen Klassifizieren im methodischen Sinn zu zählen sind, aber ohne

[4] Tatsächlich ist die Publikation von Ishikawa et al. (1994) in kaum verständlichem "Englisch" abgefasst.

[5] Details zu dieser nur in japanischer Sprache vorliegenden Arbeit sind nicht bekannt. Die obigen Angaben basieren auf dem englischsprachigen Abstract.

[6] Details zu diesem nur durch Sekundärliteratur bekannten Ansatz waren leider nicht zu ermitteln.

eigenständige Klassifizierungsleistung – und damit quasi automatisch – zur Präsenz von Notationen in den Katalogisaten führen.

(a) ***Fremddatenübernahme.*** Dies ist vermutlich die verbreitetste dieser Techniken; auf diese Weise werden etwa im *Gemeinsamen Verbundkatalog*[7] des GBV Notationen der DDC und LCC oder im *Österreichischen Verbundkatalog*[8] solche der DDC und RVK beim Erstellen von Katalogisaten aus Fremddatenbeständen (LoC, *WorldCat*, BNB, BVB) übernommen und z.T. in den OPACs weitergenutzt (etwa für weiterführende Links).

(b) ***Kooperation, Einkauf.*** Durch Zusammenarbeit oder kommerzielle Beziehungen zwischen Bibliotheken (auch Verbünden, Katalogisierungsdiensten) können durch geeignete Matching-Verfahren wie z.B. einen ISBN-Abgleich bestimmte Datenelemente (hier konkret: Notationen) aus dem jeweils anderen Katalog abgezogen bzw. importiert werden. Auf diese Weise wäre es bspw. denkbar, den *Österreichischen Verbundkatalog* mit einer grossen Zahl von UDK-Notationen aus der ETH Zürich oder Notationen der Basisklassifikation aus dem GBV auszustatten.

(c) ***Konkordanzen zwischen Klassifikationssystemen.*** Auch dadurch kann der Anteil der nach einem bestimmten System erschlossenen Bestände in Bibliothekskatalogen beträchtlich erhöht werden. Zum Thema "Konkordanzen" gibt es zahlreiche Projekte und Aktivitäten, auf die hier nicht näher eingegangen werden kann.[9] Ein Beispiel für mit Clusterverfahren erstellte Äquivalenztabellen von zwei Klassifikationen (UDK und dänische Dezimalklassifikation) im Dänischen Verbundkatalog *DANBIB* beschreibt Aagaard (1995).

7.2 Automatisches Klassifizieren von Patentliteratur

Ansätze zur automatischen Klassifikation von Patenten oder Patentanmeldungen sind m.E. als Analogiebeispiele für die automatische Erschliessung von Web-Dokumenten mit etablierten bibliothekarischen Klassifikationssystemen von besonderem Interesse. Drei Gründe sind dafür massgeblich:

(a) ***Mengen:*** Auch auf dem Patentsektor handelt es sich in der Regel um sehr grosse Datenbanken (z.B. gibt es allein 5 Millionen U.S. Patente);

(b) ***Klassifikationssysteme:*** Auch bei den Patentklassifikationen handelt es sich in der Regel um sehr grosse, hierarchische und fein verästelte Systeme;

(c) ***Dokumente:*** Wie Web-Ressourcen variieren Patentdokumente stark hinsichtlich ihrer Länge und können sehr viel Text sowie nicht-textuelle Teile (Zeichnungen etc.) enthalten. Im Vergleich zu Web-Dokumenten sind sie allerdings besser und verlässlicher in eine grosse Zahl von Abschnitten (Feldern) strukturiert. Der Text mag zwar überwiegend technisches Vokabular enthalten, weist aber auch viele legistische Terme bzw. Phrasen auf und enthält (meist um einen innovativen Eindruck zu erwecken) oft ungewöhnliche Ausdrücke; er ist damit mindestens so problematisch (heterogen, unkontrolliertes Vokabular, mitunter in verschiedenen Sprachen) wie jener von Web-Ressourcen, wenn auch weitgehend frei von grammatikalischen und typographischen Fehlern.

Mehrere Klassifikationsvorgänge in den Patentämtern würden sich für eine Automatisierung eignen. Dazu zählen das *Vorklassifizieren* ("preclassification") für das

[7] http://gso.gbv.de/ [24.06.2004]

[8] http://opac.bibvb.ac.at/acc01 [24.06.2004]

[9] Eine kurze Übersicht bietet z.B.: Einführung Nutzung DDC (2000, 40 ff.)

Weiterleiten ("routing") der neu eintreffenden Patentanmeldungen zu den Prüfern,[10] die *Reklassifizierung* von Patenten (neue Unterteilungen, Zusammenfassungen bestehender Klassen mit neuer Unterteilung, Nachziehen von Querverweisungen) sowie – wohl die schwierigste Aufgabe – das *Klassifizieren* von neuen Patenten nach dem gesamten, komplexen Klassifikationssystem (Larkey 1998; Smith 2002).

Bei der zum automatischen Klassifizieren von Patentdokumenten publizierten Literatur handelt es sich grossteils um Vorschläge, Konzepte und Tests. Nach einer neuen Quelle (Fall et al. 2003) gibt es derzeit nur ein einziges im Produktionsbetrieb verwendetes System für das automatische fachspezifische Weiterleiten von Patentanmeldungen – das weiter unten behandelte japanische System *OWAKE*.

7.2.1 Tests des U.S. Patentamtes

Bereits Ende der 1990er Jahre führte das U.S. Patentamt gemeinsam mit der Universität von Massachusetts (Amherst) ein Projekt zur Erstellung eines Retrievalsystems für Patente sowie der automatischen Klassifizierung nach der *United States Patent Classification* (USPC) durch (Larkey 1998; 1999). Bei der USPC handelt es sich um ein sehr komplexes und weitverzweigtes System mit über 400 Haupt- und rund 150.000 Subklassen auf bis zu 15 Hierarchiestufen;[11] das USPTO bildet immer neue Subklassen, die Reklassifizierungen nötig machen. Die Patente erhalten eine Hauptnotation und meist mehrere weitere Notationen (Verweisungen). Das Projekt sollte zunächst nur die Zuteilung von Haupt- und Subklasse zu neuen Anmeldungen umfassen.[12]

Larkey wählte ein Verfahren, bei dem ein neu zu klassifizierendes Dokument auf der Basis des probabilistischen Retrievalsystems *Inquery* mit allen Dokumenten einer Trainingskollektion verglichen und jener Klasse zugeteilt wurde, aus der die meisten der ähnlichsten Trainingsdokumente stammten (*k*-NN-Klassifikator). Dieses Verfahren sollte später mit einem zweiten kombiniert werden, das pro Klasse/Subklasse einen Klassifikator nach einem auf dem AIR/X-Verfahren[13] basierenden Modell trainieren würde (Naive-Bayes-Klassifikator). Durch Verwendung negativer Trainingsbeispiele sollte es besser in der Lage sein, zwischen ähnlichen Subklassen zu differenzieren.

Die Überlegungen zur Repräsentation der Dokumente sind ebenfalls interessant. Die Patente sollten nicht im Volltext verarbeitet, sondern zunächst auf bestimmte Abschnitte ("sections") bzw. Teile solcher Abschnitte reduziert werden. Der Dokumentenvektor würde dann die wichtigsten Terme bzw. Phrasen daraus enthalten, zusammen mit

[10] Dies sind z.B. im U.S. Patentamt über 3.000 Personen (Smith 2002, 270), und im Europäischen Patentamt etwa 2.500 Personen (Koster et al. 201, 20), deren fachliche Zuständigkeit den obersten Hierarchieebenen der verwendeten Patentklassifikation entsprechend aufgeteilt ist.

[11] http://www.uspto.gov/web/patents/classification/index.htm [25.07.2004]

[12] Dies betrifft immerhin rund 400.000 Fälle pro Jahr (Smith 2002, 271).

[13] Vgl. z.B.: Nohr (2003, 81 ff.)

Gewichten, die die Herkunft (Abschnitt, dem ein Term entstammt) und die Termfrequenz berücksichtigen. Konkret bedeutete dies folgende Testanordnung:

- Reduzierung des Dokuments auf Titel, Abstract, die ersten 20 Zeilen des Abschnittes "background summary", Patentansprüche;
- Entfernung aller Stoppwörter nach einer Liste des *Inquery*-Systems (418 Wörter);
- Stammformenbildung nach einem Standardverfahren;
- Jeder Wortstamm, der mindestens zweimal auftrat, galt als potentielle Vektorkomponente;
- Bildung der Gewichte nach Abschnitten (dabei erhielt der Titel dreimal soviel Gewicht wie die anderen Abschnitte) und Häufigkeit der Terme im betreffenden Abschnitt;
- Bildung der Summe der Gewichte pro Term (über alle Abschnitte);
- Rangreihung nach diesen Summen und Anwendung eines Schwellenwertes (maximal *n* Terme pro Vektor; Mindestgewicht = 2);
- Analoge Vorgangsweise für die mittels eines Parsers ("part of speech tagger") extrahierten Phrasen. Die Einbeziehung von Phrasen wurde jedoch wieder fallengelassen, da sich damit keine besseren Ergebnisse erzielen liessen.

Die Testergebnisse selbst wurden leider nicht veröffentlicht, was darauf hindeutet, dass wohl keine für einen Routinebetrieb zufriedenstellenden Resultate erzielt werden konnten. Aus einer neueren Publikation (Smith 2002) geht hervor, dass das USPTO bislang nur eine Retrievalkomponente verwendet, die zu einem gegebenen Dokument eine ranggeordnete Liste der ähnlichsten Patente in der Datenbank sowie deren Notationen ausgibt. Dies wird von Patentprüfern und Klassifizierern als Unterstützung bei der Entscheidung bezüglich der zu vergebenden Notationen genutzt.

7.2.2 Tests des Europäischen Patentamtes

Auch im EPA ist eine automatische Vorklassifizierung der einlangenden Patentanmeldungen[14] zum Zweck des Weiterleitens an die fachlich zuständigen Prüfer von vorrangigem Interesse. Tests, die 1999 und 2000 mit einer Reihe von Firmen und Institutionen durchgeführt wurden, sollten helfen, ein geeignetes Verfahren dafür zu finden (Krier & Zaccà 2002; Koster et al. 2001; 2003). Als Grundanforderung war vorgegeben, die langjährig bekannte Güte der intellektuellen Vorklassifikation bei der Kategorisierung nach den 44 Prüfungsdirektionen zu erreichen (81,2%); nach Möglichkeit sollte auch noch genauer klassifiziert werden können, entweder nach den insgesamt 549 fachlichen Prüferteams oder nach 624 Subklassen.

Im folgenden soll nur auf den zweiten, umfangreicheren Test eingegangen werden. Dabei erhielten die teilnehmenden Institutionen – jeweils in maschinenlesbarer Form – für jede der 44 Kategorien 2.000 klassifizierte Trainingsdokumente sowie 1.000 "anonyme" Testdokumente, d.h. solche, bei denen die richtigen Notationen nur dem

[14] Jährlich etwa 140.000 (Koster et al. 2001, 20).

EPA bekannt waren. Obwohl das EPA Anmeldungen in englischer, deutscher und französischer Sprache erhält, wurden für den Test nur englischsprachige Anmeldungen verwendet. Die Aufgabe der Teilnehmer bestand darin, für jedes Testdokument die am besten zutreffende Kategorie zu liefern, zusammen mit einem Koeffizienten, der die Sicherheit dieser Zuteilung ausdrücken sollte. Nach Möglichkeit sollte auch noch eine zweitbeste Notation geliefert werden. Die Tests waren sowohl für die Volltexte der Anmeldungen als auch nur für die Abstracts durchzuführen.

Die Auswertung der Testergebnisse erfolgte sodann im EPA. Dabei war Precision (wie üblich) als Übereinstimmung der gelieferten Notation mit einer der bei der früher durchgeführten intellektuellen Klassifizierung zugewiesenen Notationen definiert. Als Recall wurde der Anteil der bis zu einem bestimmten Schwellenwert sicher klassifizierten Dokumente gewertet, da man davon ausging, dass es auch in der Praxis besser sein würde, die vom Verfahren unsicher zugeordneten Dokumente manuell zu bearbeiten. Interessante Erfahrungen wurden hinsichtlich der *Geschwindigkeit* gemacht: Pro Volltext wurden für das Klassifizieren zwischen 0,5 und 4 Sekunden benötigt, was eine akzeptable Dauer für die tägliche Produktion erwarten liess (maximal 30 Minuten); für die Trainingsphase wurde bis zu einer Woche (ohne Unterbrechung) benötigt. Im Vergleich zur Verarbeitung nur der Abstracts erbrachte die Verwendung der Volltexte zwischen 2% und 9% höhere Precision.

Die Ergebnisse blieben jedoch hinter jenen der manuellen Bearbeitung zurück. Das am besten gereihte Verfahren erzielte bei 100% Recall auf der Ebene der 44 Prüfungsdirektionen nur 72% Precision und auf der Ebene der 459 Teams nur mehr 57%. Um den bisherigen Gütewert von 81,2% (oberste Ebene) zu halten, könnten nur 78% der einlangenden Anmeldungen automatisch weitergeleitet werden. Die unbefriedigenden Resultate auf der feineren Ebene wurden z.T. darauf zurückgeführt, dass für diese Klassen zu wenige Trainingsdokumente verfügbar und diese auch nicht gleichmässig auf die Subklassen verteilt gewesen waren.

Im EPA entschloss man sich dennoch, einer schrittweisen Implementierung näherzutreten und neue Versuche zur Erreichung einer besseren Klassifizierungsgüte auf dem feineren Niveau durchzuführen. Nach EPA (2004, 20) findet die automatische Vorklassifizierung in einigen technischen Gebieten bereits Anwendung. Als weiteres Anwendungsfeld wurde auch eine bislang in klassifikatorischer Hinsicht vernachlässigte Literaturdokumentation genannt, für die eine *grobe* automatische Klassifizierung in jedem Fall besser wäre als überhaupt keine.

Nähere Details zu dem teilnehmenden Verfahren mit den relativ besten Resultaten (Universität Nijmegen, Niederlande) enthalten die Arbeiten von Koster et al. (2001, 2003). Danach wurde ein Balanced-Winnow-Algorithmus eingesetzt, ein zur Kategorie der inkrementellen Klassifikatoren zählendes heuristisches Lernverfahren (vgl. Sebasti-

ani 2002b, 25), dieses gegen einen gebräuchlichen Rocchio-Algorithmus getestet und dabei für besser befunden. Die Dokumente wurden ohne linguistische Vorbereitung verarbeitet (keine Eliminierung von Stoppwörtern, kein Stemming), die interne Struktur der Dokumente (Abschnitte, Felder) wurde im Gegensatz zu dem oben beschriebenen Verfahren von Larkey (1998; 1999) völlig ignoriert.

7.2.3 Tests der WIPO

Die von der WIPO in englischer und französischer Sprache herausgegebene[15] und auch in zahlreichen anderen Sprachen vorliegende *Internationale Patentklassifikation* (IPC) ist das am weitesten verbreitete Klassifikationssystem für Patentliteratur. Die IPC ist ein komplexes, hierarchisches und sehr detailliertes System, das sich in acht Sektionen (z.B. "A"), 120 Klassen ("A61"), etwa 650 Unterklassen ("A61K") und rund 69.000 Haupt- und Untergruppen ("A61K 9/00", "A61K 9/06") gliedert. Von der WIPO selbst wird sie u.a. für die Klassifizierung der unter dem *Patent Cooperation Treaty* eingereichten Patentanmeldungen verwendet.

Vor kurzem wurden für die WIPO Tests zur automatischen Klassifizierung nach der IPC durchgeführt (Fall et al. 2003). Für diese Untersuchung wurde die neu erstellte Testdatenbank *WIPO-alpha*, eine auch öffentlich verfügbare Kollektion OCR-konvertierter, englischsprachiger Patentanmeldungen aus verschiedenen Ländern,[16] verwendet. Diese Datenbank umfasst ca. 75.000 Dokumente aus dem Zeitraum 1998 bis 2000 und besteht aus einer Trainingskollektion (46.000 Dokumente) und einer Testkollektion (29.000). Die Dokumente enthalten die mit XML-Tags versehenen Felder Titel, Erfinder, Einreichende Firmen/Personen, Abstract, Patentansprüche und Langbeschreibung, sowie (immer) eine IPC-Hauptnotation und (oft) IPC-Sekundärnotationen. Die Notationen sind auf der Ebene der Klassen ("D01") sowie der Unterklassen ("D01D") enthalten; die Dokumente sind so "natürlich" wie möglich über die Hierarchien verteilt, wobei die sehr gering besetzten Klassen bzw. Unterklassen nicht vertreten sind.[17]

Für die Tests wurden vier verschiedene Klassifikatoren (Naive Bayes, *k*-NN, SVM und Winnow) herangezogen und verglichen. Die angewandten Aufbereitungsstrategien inkludierten die Verwendung bzw. Nichtverwendung von Stoppwörtern und Stemming, die Variation von Wort- bzw. Dokumentenhäufigkeit bei der Gewichtung und die Klassifizierung auf Basis verschiedener Datenfelder (Titel, Patentansprüche, die ersten 300 Wörter aus allen Feldern in der o.a. Reihenfolge). Auf die Details der diversen Anordnungen kann hier ebensowenig eingegangen werden wie auf die Einzelergebnisse für die verschiedenen Klassifikatoren. Interessant sind aber die drei als Masse für

[15] http://www.wipo.int/classifications/ [25.06.2004]

[16] http://www.wipo.int/ibis/datasets/ [25.06.2004]

[17] Die Kollektion enthält Dokumente zu 114 Klassen und 451 Unterklassen.

die Precision verwendeten Kriterien, mit denen der Output der Klassifikatoren, d.h. die pro Dokument ranggeordnete Liste der Notationen für die Klassen bzw. Unterklassen, untersucht wurde (*Abbildung 7-1*):[18]

(a) ***Top prediction:*** Vergleich der bestgereihten Notation mit der realen IPC-Hauptnotation;

(a) ***Three guesses:*** Vergleich der drei bestgereihten Notationen mit der realen IPC-Hauptnotation;

(b) ***All categories:*** Vergleich der bestgereihten Notation mit allen im Dokument real enthaltenen Notationen.

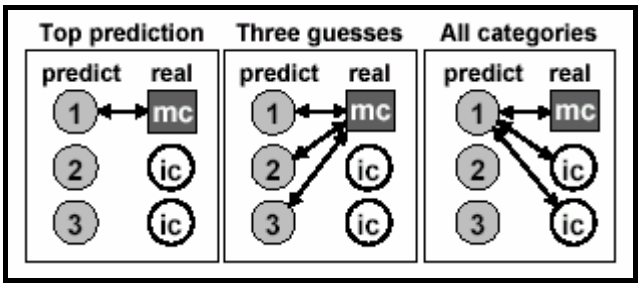

Abbildung 7-1: WIPO-Tests – Drei Gütekriterien (Quelle: Fall et al. 2003, Kpt. 3)

Diese Masse wurden als Indikatoren dafür gewertet, ob sich das Verfahren als autonomer Klassifikator oder zur Unterstützung eines menschlichen Klassifizierers eignen würde – für den ersteren Zwecke müsste ein guter Wert für "top prediction" erzielt werden, während für den zweiten die weniger strikten Bedingungen ausreichen könnten.

Die Ergebnisse für die Zuteilung zu den 114 Klassen zeigten, dass dafür der Ansatz mit den ersten 300 Wörtern der Dokumente am besten abschnitt und die Patentansprüche allein die schlechteste Basis für die automatische Klassifizierung darstellten. Allerdings erreichte keines der getesteten Verfahren mehr als 55% korrekter Übereinstimmungen der bestgereihten Notation mit der realen Hauptnotation, während beim "three guesses"-Vergleich Werte zwischen 73% und 79% resultierten.

Die Resultate für die 451 Unterklassen erbrachten hinsichtlich der exakten Übereinstimmung noch schlechtere Werte – den relativ besten (41%) vermochte der SVM-Klassifikator beim Ansatz mit den ersten 300 Wörtern zu erzielen. Ein weiteres Detailergebnis deutete darauf hin, dass diese 41% durch das Trainieren des Klassifikators *nur* mit solchen Dokumenten, die nicht mehr als *eine* Unterklassen-Notation aufweisen, noch verbessert werden könnte. Dazu müsste jedoch *WIPO-alpha* erheblich vergrössert werden.

Die Schlussfolgerung aus diesen Ergebnissen war: "..., the combinations of algorithms and training collections reported here do not appear sufficient for building a fully-automated system for the categorization of all patent applications." (Fall et al.

[18] "mc" = Hauptnotation, "ic" = sekundäre Notation

2003, Kpt. 5). Allerdings könne man sich sehr wohl ein semi-automatisches System vorstellen, das dem menschlichen Klassifizierer eine Zahl von Notationen vorschlagen würde. Ein solches System sollte aber mehr als "three guesses" präsentieren, da bei einer etwas grösseren Zahl von Notationen die Wahrscheinlichkeit, darunter die richtige zu finden, wesentlich besser sei. Die WIPO beabsichtigt die baldige Realisierung eines solchen halbautomatischen Hilfsmittels.

7.2.4 Die französische IPC-Suchmaschine

Eine Anwendung, die die IPC zur Grundlage hatte, war auch der in Arbeiten von Leclercq (1999) und Lyon (1999) vorgestellte Dienst *CIB-LN* des *Institut National de la Propriété Industrielle* (INPI). Dabei handelte es sich aber nicht um ein Klassifizierungsverfahren im engeren Sinn, sondern um eine Suchmaschine, die interessierten Endbenutzern einen natürlichsprachlichen Zugriff (in französischer Sprache) zur IPC (französisch: "CIB") ermöglichen sollte, um damit die Recherche der frei zugänglichen Patentdatenbank *Web Brevets* bzw. anderen nach der IPC erschlossenen Datenbanken zu erleichtern. Dabei wurden Notationen bis zur Hauptgruppenebene der IPC ausgegeben (die Benennungen der Untergruppen waren den zugehörigen Hauptgruppen zugeteilt worden).

Aus den zitierten Literaturstellen geht hervor, dass *CIB-LN* 1998 für das INPI von einer Softwarefirma auf der Basis zweier kommerzieller Produkte – des Sprachverarbeitungsprogramms *LexiQuest* und der Suchmaschine *Verity IS97* – erstellt wurde. Durch die Kombination dieser Programme wurde folgendes Leistungsspektrum erzielt:

- Linguistische Analyse der Anfrage:
  - Morphologische Analyse v.a. zur Worterkennung (konjugierte bzw. flektierte Wortformen, Komposita);
  - Syntaktische Analyse zur Ermittlung und Bewertung von Mehrwortgruppen;
  - Semantische Analyse zur Disambiguierung und Ergänzung von Synonymen;
  - Erstellung vereinfachter, gewichteter logischer Repräsentationen der Anfrage und Übersetzung derselben in die Syntax der Suchmaschine.
- Selektion/Ausgabe der IPC-Notationen:
  - Die IPC-Datenbank basierte auf dem aus Registern usw. angereicherten Vokabular der IPC;
  - Die gefundenen Notationen wurden als ranggeordnete Liste ausgegeben, die mit Links versehen war, über welche die volle Information zur jeweiligen Klasse, Unterklasse oder Gruppe abgerufen bzw. dann in den Hierarchien der IPC weiternavigiert werden konnte.

Ein 1999 durchgeführter Test des Systems inkludierte 350 Endbenutzeranfragen, die von IPC-Experten mit allen passenden Notationen versehen und danach mit den aus *CIB-LN* stammenden Resultaten verglichen wurden. Dabei zeigte sich, dass das System

79% der von den Experten vergebenen Notationen fand und 55% der manuell vergebenen Notationen unter den erstgereihten 20 aus *CIB-LN* aufschienen (Lyon 1999, 94).

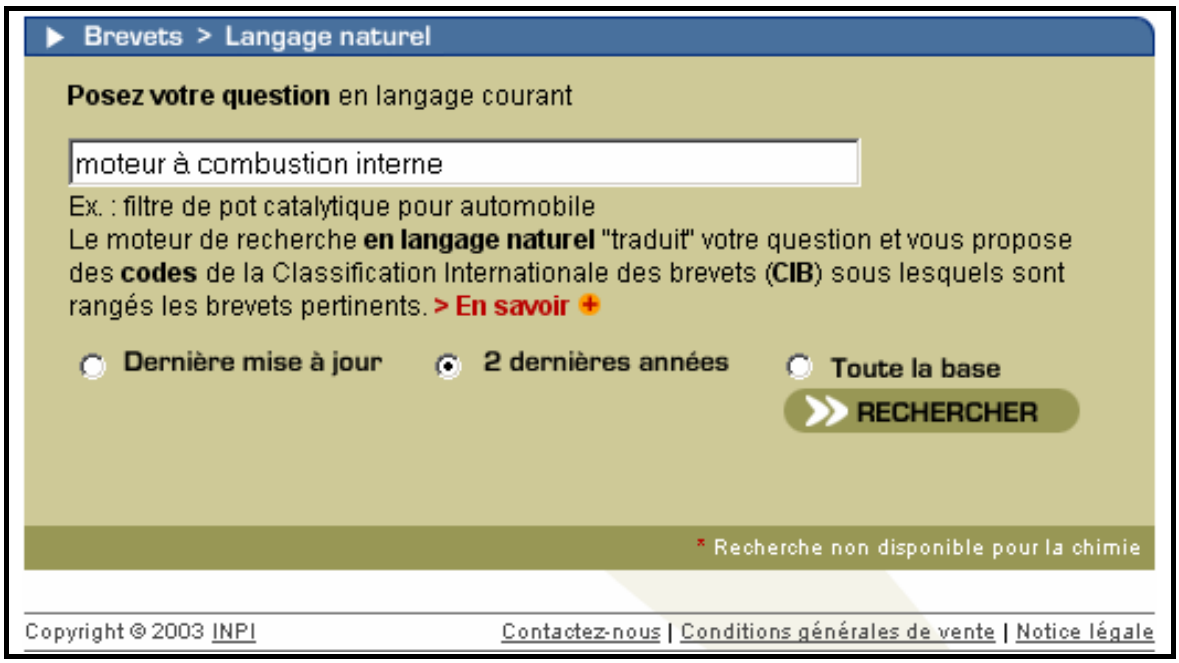

Abbildung 7-2: *Plutarque* – Natürlichsprachliche Suche

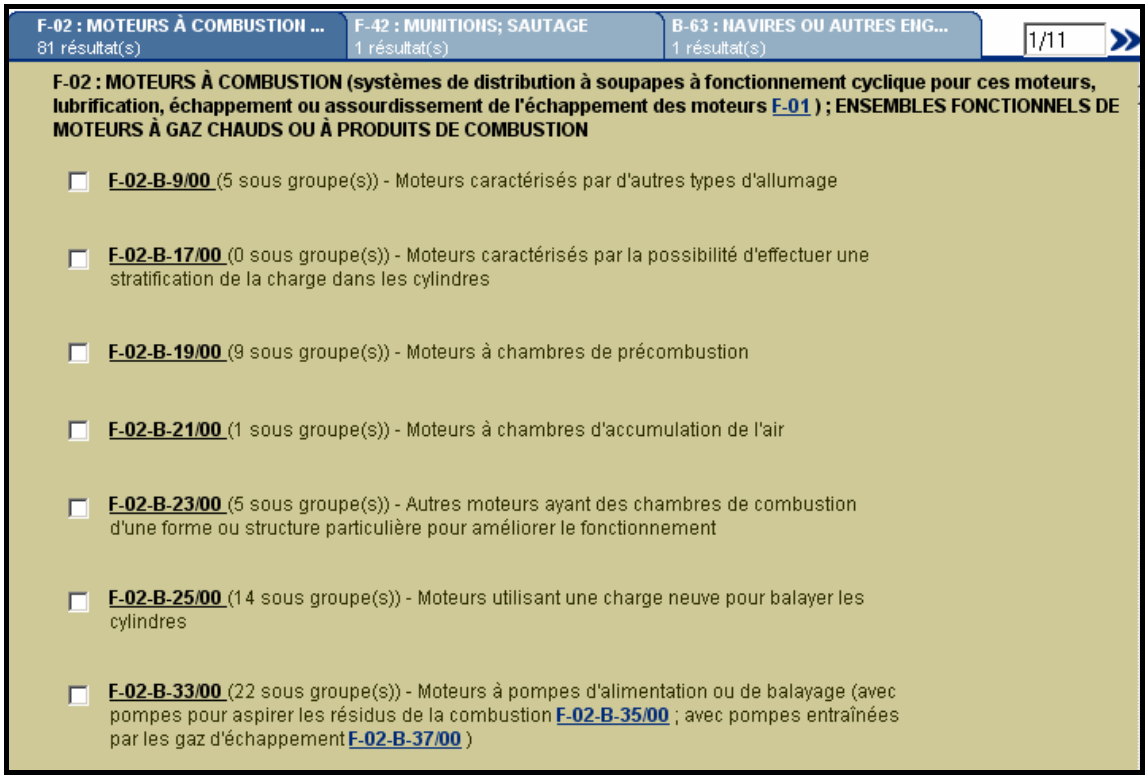

Abbildung 7-3: *Plutarque* – Anzeige der IPC

Aktuelle Recherchen ergaben, dass *CIB-LN* und *Web Brevets* heute nicht mehr angeboten werden. Die in der Literatur beschriebenen Funktionalitäten liessen sich jedoch in *Plutarque* aufspüren,[19] einer Patentdatenbank, die auch nicht registrierten Benutzern für kurze Zeiträume frei zur Verfügung steht. Eine der Suchoptionen in *Plutarque* ist die Recherche in natürlicher Sprache, bei der die Eingabewörter in Notationen der IPC "übersetzt" werden (*Abbildung 7-2*). Dies führt zur Anzeige des betreffenden Ausschnittes aus der IPC auf Hauptgruppenebene (*Abbildung 7-3*) mit der Option, allenfalls auch noch zu den Untergruppen zu navigieren bzw. zur Anzeige der entsprechenden Patente aufgrund der Auswahl der gewünschten Klassen durch den Benutzer.

7.2.5 Das japanische Klassifizierungssystem OWAKE

Wie eingangs erwähnt, ist *OWAKE* das bislang einzige im Produktionsbetrieb verwendete automatische Verfahren zur (Grob-)Klassifizierung[20] von Patentanmeldungen. Dieses erst kürzlich in Europa vorgestellte System (Shimizu 2003) wurde 2002 vom japanischen *Industrial Property Cooperation Center* (IPCC) entwickelt, das für die Klassifizierung neuer Patentanmeldungen[21] mittels der japanischen Patentcodes ("F-terms") und der IPC zuständig ist.

Während die IPC-Notationen nach wie vor nur intellektuell vergeben werden, soll *OWAKE* die Indexierer bei der Vergabe der japanischen Patentcodes unterstützen. Bei diesen "F-terms" handelt es sich um etwa 3.000 Codes für technische Anwendungsfelder wie z.B. "Cooking vessels" (4B055), "Semiconductor devices" (4M001) oder "Polymer compositions" (4J002), die 38 übergeordneten Gruppen (z.B. "3L" für "Heat machinery") zugeordnet sind. Das Arbeitsprinzip des Verfahrens besteht im Vergleich des Vokabulars einer neuen Anmeldung mit einer Datenbank aus dem Vokabular publizierter Patente, wodurch – über den "Umweg" der mit diesem Vokabular assoziierten groben IPC-Notationen (auf der Ebene der Klassen) die "F-terms" der passenden Patentpublikationen ermittelt und gereiht werden. *Abbildung 7-4* veranschaulicht diesen Prozess etwa detaillierter.

Für den Aufbau der Datenbank wird das Vokabular der jeweils aktuellsten Patentschriften (CD-ROM mit ca. 380.000 Datensätzen) herangezogen, nach Abschnitten (Titel, Abstract, Ansprüche etc.) getrennt aufbereitet und mit Häufigkeitsgewichtungen (ganze Datenbank, einzelne Klassen der IPC) versehen. Bei der Verarbeitung eines neu zu klassifizierenden Dokuments wird dessen Vokabular in einem ersten Schritt mit dem IPC-Klassen-spezifischen Vokabular der Datenbank verglichen, wodurch die drei ähn-

[19] http://www. plutarque.com [03.07.2004]

[20] Nach Shimizu (2003) bedeutet OWAKE (大分け) "big-scale classification"; gemeint ist wohl eine nicht sehr feine Klassifizierung.

[21] Pro Jahr ca. 400.000 Anmeldungen.

lichsten IPC-Klassen ermittelt werden. Im zweiten Schritt erfolgt sodann der Vergleich des Vokabulars aller Patentschriften, die zu diesen drei Klassen gehören, mit dem Vokabular des Eingabedokuments, wodurch jene 50 Patentschriften gefunden werden, die dem Dokument am ähnlichsten sind (*k*-NN Verfahren). Für jeden in diesen 50 Dokumenten enthaltenen "F-term" (in der Abbildung "theme codes" genannt) werden die Ähnlichkeitskoeffizienten der betreffenden Dokumente aufsummiert (wobei sekundäre "F-terms" oder "sub theme codes" nur zur Hälfte zählen). Der thematische Code, der dabei die höchste Summe erzielt, wird der Patentanmeldung als *OWAKE*-Vorschlag zugeteilt.

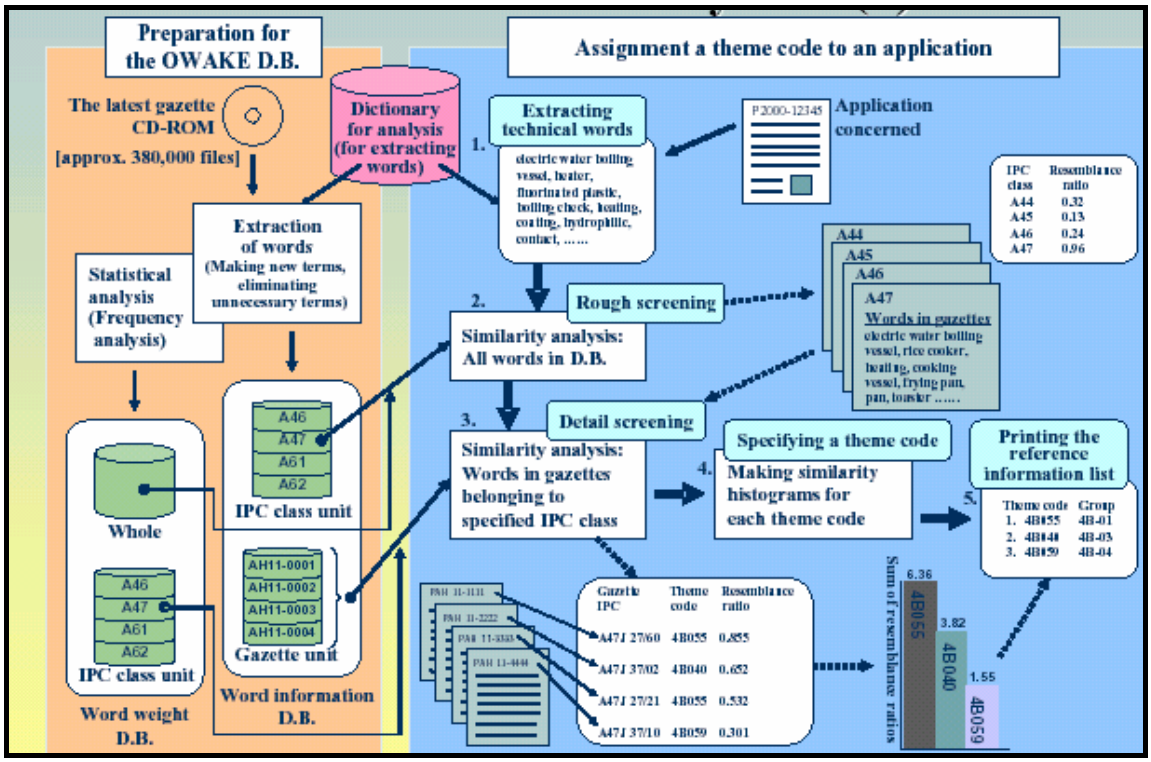

Abbildung 7-4: *OWAKE* – Klassifizierungsprozess
(Quelle: Shimizu 2003)

Shimizu (2003) berichtet von durchschnittlich 60% korrekten Zuteilungen auf der Ebene der "F-terms" und 90% auf jener der übergeordneten, wesentlich allgemeineren "groups". Bei letzteren wird jedoch eine themenabhängige Schwankungsbreite von 95% ("Business machinery group") bis 75% ("Digital Communication group") angegeben, sodass zu vermuten ist, dass die Variation auf der Ebene der "F-terms" wohl noch grösser sein dürfte.

Das IPCC arbeitet aber bereits an dem Ablösesystem *CYUWAKE*,[22] das schon Ende 2004 auf der Ebene der "groups" 95% und auf der Ebene der "F-terms" 90% kor-

[22] Dieser Name (中分け) bedeutet "medium-scale classification" (ibid.); gemeint ist wohl eine Klassifizierung auf etwas feinerem Hierarchieniveau.

rekte Zuordnungen erreichen soll. Ausserdem strebt man damit auch die Ermittlung der IPC-Hauptnotation auf Untergruppenebene mit einer Güte von 60% bis 67,5% an. Der Hauptanteil dieser Verbesserungen soll durch eine Vergrösserung und strukturelle Optimierung der Datenbank erzielt werden.

7.3 Automatisches Klassifizieren in der Mediendokumentation

Auch Artikel aus Zeitungen, Zeitschriften und Magazinen sowie Presse- und Agenturmeldungen sind als Objekte einer automatischen Klassifizierung von Interesse, da sie ebenfalls Ähnlichkeit mit bestimmten Web-Dokumenten aufweisen (thematische Vielfalt, unterschiedliche Länge, geringer Strukturiertheitsgrad, unkontrolliertes Vokabular). Im Gegensatz zu Patentschriften entstammt ihr Vokabular aber nicht der Fachsprache, sondern dem alltäglichen Gebrauch – wie auch bei einem Teil der via WWW zugänglichen Ressourcen. Im Hinblick auf die verwendeten Klassifikationsschemata besteht jedoch ein grosser Unterschied zu den aus dem Bibliotheks- bzw. aus der Patentbereich kommenden, da es sich dabei in der Regel um keine grossen, komplexen Systeme handelt. Daher sollen die folgenden Anwendungen und Projekte auch etwas weniger ausführlich behandelt werden.

7.3.1 Gruner + Jahr

Seit 1998 setzt das Verlagshaus *Gruner + Jahr* (G+J) ein Verfahren zur (semi-)automatischen Textklassifizierung bei der Erschliessung seiner Pressedatenbank ein – zunächst in einem Testbetrieb und seit Anfang 2002 im Produktionsbetrieb (Gaese 2003). "Klassifizieren" bedeutet im Kontext von G+J die Zuteilung von Deskriptoren aus einem hierarchischen Thesaurus von etwa 1.300 Begriffen. Das Verfahren, *DocCat* (Document Categorizer), stammt ursprünglich von IBM, wo es speziell für G+J entwickelt wurde.[23] Die Klassifizierungskomponente, die als linguistisches Modul nur eine Lemmatisierungsfunktion verwendet, basiert auf dem Lernen eines repräsentativen, klassifizierten Trainingskorpus – bei G+J waren hierfür sowie für Tests rund 2 Millionen gespeicherter Volltexte von Presseartikeln vorhanden. Technische Details zu dem eingesetzten Klassifikator sind nicht bekannt. Neben der Klassifizierung extrahiert das System auch freie Schlagwörter und erkennt Personen- und Organisationsnamen, wofür weitere Techniken wie wörterbuchgestützte morphologische Analyse, Namensgrammatik, Wissensbank, hinterlegte Listen von Vornamen etc. Anwendung finden.

[23] 2000 gingen die Rechte für die Software auf die Firma TEMIS S.A. über, die seither das Programm betreut und weiterentwickelt.

Die Ergebnisse aus dem Testzeitraum (1999–2001), in dem das System immer wieder verbessert und mit einer grossen Zahl verschiedener Klassifikatoren getestet wurde, zeigten, dass mit *DocCat* folgendes erreicht werden kann:

- die tägliche Produktion von ca. 1.100 neu hinzukommenden Presseartikeln aus unterschiedlichen Medientypen in knapp 20 Minuten;
- ein Recall zwischen 60% und 80% (Anteil, zu dem das System jene Deskriptoren verwendet, die bei der parallel durchgeführten intellektuellen Erschliessung vergeben wurden);
- eine Precision zwischen 75% und 90% (Anteil der Übereinstimmung zwischen maschineller und intellektueller Zuteilung).

Recall und Precision variierten in Abhängigkeit von Inhalt und Stil; "feuilletonistische" Themen sowie "literarisch" geschriebene Artikel wurden weniger gut klassifiziert. Die ursprünglich geplante automatische Zuteilung von Personendeskriptoren aus einem speziellen Thesaurus sowie von Ländercodes wurde wegen nicht zufriedenstellender Testergebnisse nicht realisiert (vgl. auch Rapke 2001). Die künftige Verwendbarkeit des Verfahrens für englisch- und französischsprachige Dokumente wird erwartet.

Die gegenwärtige Version von *DocCat* wird täglich für die Erschliessung aller deutschsprachigen Volltexte von Presseartikeln eingesetzt. Sie arbeitet zweistufig: Im ersten Schritt werden pro Dokument bis zu vier Themenbereiche zugeteilt; im zweiten Schritt vergeben spezifische Klassifikatoren für diese Themenbereich jeweils zwei Deskriptoren. Diese dienen den menschlichen Indexierern als Vorschläge, die akzeptiert, verändert oder verworfen werden können.

7.3.2 Zweites Deutsches Fernsehen

Kleinoeder & Puzicha (2002) berichteten über einen Pilottest, der vom ZDF zur Klassifizierung von Pressemeldungen (Zeitungs-/Zeitschriftenartikel, Agenturberichte usw.) mit einem kommerziellen Produkt – *MindServer Categorization* der Firma *Recommind*[24] – durchgeführt wurde. Diese Software arbeitet mit dem statistischen Verfahren "Probabilistic Latent Semantic Analysis" (PLSA), das durch eine Häufigkeitsanalyse gemeinsam auftretender Wörter im Textkorpus bedeutungstragende Wortgruppen identifiziert die Dokumente auf dieser Basis clusteranalytisch kategorisiert oder vorgegebenen "Taxonomien" zuordnet. Sie nimmt für sich in Anspruch, Mehrdeutigkeit aufgrund der Erkennung multipler Kontexte, in denen die betreffenden Wörter verwendet werden, reduzieren zu können und darüberhinaus sprachenunabhängig zu sein (Recommind o.J.)

Dem Test lagen 15.586 nach einem ZDF-eigenen Schema (167 Kategorien auf fünf Hierarchiestufen) manuell erschlossene Dokumente zugrunde. Von diesen wurden nach dem Zufallsprinzip 80% als Trainings- und 20% als Testkollektion ausgewählt. Die Untersuchung der Klassifizierungsergebnisse erbrachte eine durchschnittliche Pre-

[24] http://www.recommind.com [06.04.2004]

cision von 75,3% (korrekte Zuordnungen) und einen Recall von 91,6% (vollständige Zuordnungen zu allen menschlich vergebenen Klassen). Die Autoren bewerteten diese Resultate zwar als positiv, regten aber dennoch eine "semiautomatische Integration" des Verfahrens, d.h. seine Verwendung zur Unterstützung menschlicher Klassifizierer, an.

7.3.3 Bayerischer Rundfunk und Süddeutscher Verlag

Eine Machbarkeitsstudie für die fast vollständige Automatisierung des Klassifizierungsprozesses in einem neuen elektronischen Pressearchiv[25] wurde im Rahmen einer am Institut für Rundfunktechnik, München, durchgeführten Diplomarbeit erstellt (Winkler 1997). Dabei ging es um die Archivierung von täglich rund 1.400 Artikeln aus Zeitungen und anderen Printmedien in virtuellen "Mappen" (jeweils einer oder mehreren davon), die nach einer bestehenden Systematik mit vier bis fünf Hierarchieebenen geordnet waren. Als Ziel der automatischen Klassifizierung wurde die Erreichung von 100% korrekten Zuordnungen (wohl zulasten der Klassifizierbarkeit aller Dokumente) ins Auge gefasst; nicht eindeutig klassifizierbare Texte müssten weiterhin manuell bearbeitet werden.

Das entworfene System sah eine nicht ganz unkomplizierte Anordnung vor, bei der im ersten Schritt eine automatische Indexierung der Dokumente mit Hilfe eines aus den Klassenbenennungen und zusätzlichem Vokabular erstellten Thesaurus erfolgen sollte. Als einziges linguistisches Verfahren sollte dabei ein Algorithmus zur Erkennung von Mehrwortgruppen eingesetzt werden. Im zweiten Schritt sollte dann die Klassifizierung – nur mehr auf der Basis dieses kontrollierten Vokabulars – erfolgen, wofür die Kombination eines Clusterverfahrens mit einer ganzen Reihe verschiedener heuristischer Regeln geplant wurde.

Dieser stark auf die Nachbildung der manuellen Klassifizierung und der Einbeziehung vieler daraus abgeleiteter Bearbeitungsregeln abzielende Entwurf wirkt zwar praxisorientiert, gleichzeitig aber eher umständlich und auch methodisch wenig modern.

Eine aktuelle Recherche erbrachte, dass das Dokumentations- und Informationszentrum München des Süddeutschen Verlags heute tatsächlich ein automatisches Klassifizierungsverfahren einsetzt und auch als Dienstleistung anbietet.[26] Dieses hat offensichtlich mit dem oben beschriebenen Entwurf nichts zu tun; vielmehr wurde es in Kooperation mit zwei Softwarefirmen entwickelt. Es wird vor allem für die halbautomatische Klassifizierung eingesetzt (Erstellung von Klassifizierungsvorschlägen), kann aber in einigen Teilbereichen sogar vollautomatisch arbeiten. Aus der Firmenliteratur

[25] Damals geplante Zusammenfassung der beiden getrennt bestehenden, bislang nur manuell bearbeiteten Textarchive des Bayerischen Rundfunks und des Süddeutschen Verlages.

[26] Vgl.: http://www.diz-muenchen.de/html/autoklassifizierung.html [28.06.2004]

(Hees 2004, 12) geht hervor, dass es sich um ein am Beispiel von Trainingsdokumenten lernendes Verfahren handelt.

7.3.4 Artikel aus belgischen Magazinen

Von einer an der Universität Löwen für einen belgischen Verlag durchgeführten Studie zur automatischen Klassifizierung von Artikeln aus Publikumszeitschriften in flämischer Sprache berichten Moens & Dumortier (1999; 2000). Die Anforderung bestand in der Zuteilung zu einem groben Schema thematischer Kategorien zum Zweck der Verteilung der elektronischen Versionen dieser Artikel an bestimmte Gruppen von Abonnenten. Für die Studie wurden 14 solcher Kategorien (z.B. "car", "investments", "film", "tourism") ausgewählt. Für den Test standen 2.650 bereits früher manuell kategorisierte Artikel, die in der Länge stark variierten und meist nur 1 oder 2 Kategoriezuordnungen aufwiesen, zur Verfügung. Zwei Drittel davon wurden als Trainings- und ein Drittel als Testkollektion genutzt. Als Attribute wurden aus den Dokumenten jeweils Eigennamen und signifikante Wörter extrahiert, wobei sich für die Selektion der letzteren eine umfangreiche Stoppwortliste und eine Termgewichtung[27] als effizienter herausstellten als etwa die Auswahl von Wörtern aus dem Anfangsteil der Dokumente. Auf Stemming wurde generell verzichtet. Mit dem besten von drei getesteten Verfahren – einem auf dem bekannten χ^2-Test (der manchmal für die Attributauswahl verwendet wird) basierenden Klassifikator – konnten durchschnittlich 73% Recall und 64% Precision (Zuordnung zu ein bis zwei Kategorien) bzw. 69% Recall und 68% Precision (Zuordnung zu nur einer Kategorie) erzielt werden. Dabei war allerdings eine beträchtliche Variation zwischen den Kategorien zu registrieren (z.B. erreichte "Marketing" 91% Precision, "Politik" dagegen nur 38%). Das Ergebnis wurde als zufriedenstellend erachtet und führte zur Aufnahme des Echtbetriebes mit diesem Verfahren.

7.3.5 Andere Untersuchungen an Pressetexten

Im Gegensatz zu der eben referierten Studie dienten die meisten der zahlreichen Untersuchungen zum automatischen Klassifizieren, die anhand von Pressemeldungen, Newsfeeds u.dgl. vorgenommen wurden, nicht der Erstellung eines Verfahrens für den praktischen Einsatz, sondern der akademischen Arbeit, insbesondere auf dem methodischen Sektor (vgl. Abschnitt 2.6.2). Dies trifft insbesondere auf die Studien mit der bekannten Testkollektion *Reuters-21578* zu, die Meldungen von Nachrichtenagenturen enthält, die nach 135 Kategorien mit Wirtschafts- und Politikbezug klassifiziert sind. Auch ein weiteres studentisches Projekt zur Kategorisierung von ca. 23.000 Zeitungsartikeln (Müller 2002) zählt zu dieser Klasse von Arbeiten.

[27] Termfrequenz, normalisiert mittels der inversen Frequenz des im Dokument am häufigsten auftretenden Terms.

Beispiele für Untersuchungen an Pressekollektionen aus anderen Sprachräumen sind die Studien von Goller et al. (2000) bzw. Figuerola et al. (2001), in denen Artikel aus der *Süddeutschen Zeitung* bzw. aus *El Mundo* verwendet wurden. Auch in diesen Arbeiten ging es nur um methodische Fragen, auf die hier nicht weiter eingegangen werden soll.

7.4 Anwendungen bei Web-Portalen, Suchmaschinen, Informationsdiensten

7.4.1 Lexis-Nexis

Der bekannte kommerzielle Dienst *Lexis-Nexis,*[28] der seit 1973 zunächst legistische Informationen und seit geraumer Zeit auch Nachrichten sowie Informationen aus Wirtschaft und öffentlicher Verwaltung anbietet, wendet bereits seit den frühen 1990er Jahren ein Verfahren zur automatischen Dokumentenerschliessung an (Schmeer & Sidlo 1998). Dabei wird "automatisches Klassifizieren" eigentlich im Sinne des automatischen Indexierens verstanden, da es dabei darum geht, natürlichsprachliche Volltexte mit Begriffen aus einem kontrollierten Vokabular zu versehen. Das eingesetzte Verfahren wurde ursprünglich als *Term-based Topic Identification* bezeichnet und nach weiteren Modifikationen unter den Benennungen *NEXIS®Indexing, SmartIndexing* und *Topical Indexing* vermarktet. Neben Deskriptoren verschiedener Art (Fachgebiete, Firmen- und Personennamen usw.) sollen dabei auch SIC-Notationen[29] vergeben worden sein (ibid., 342).

Ende der 1990er Jahre betrug die Zahl der bei *Lexis-Nexis* gespeicherten Dokumente bereits über eine Milliarde (ibid., 338). Die enormen quantitativen Ausmasse hatten offensichtlich entscheidend zur Einführung der automatischen Erschliessung beigetragen. Das entwickelte Verfahren geht auf eine ursprünglich studentische Arbeit aus den 1980er Jahren zurück, die später modifiziert wurde und schliesslich für eine nahezu vollautomatische Zuordnung der Dokumente zu über 70.000 Kategorien (Firmen, Personen, Institutionen, Orte, Themen aus aktuellem Geschehen, Wirtschaft und Verwaltung) sorgte (Wasson 2001). Die Dokumentenkollektion setzt sich aus Nachrichtenmeldungen, Firmenberichten, legistischen Publikationen und Web-Ressourcen zusammen.

Bei der eingesetzten Technik handelt es sich, soweit dies der verfügbaren Literatur entnommen werden kann, um ein regelbasiertes Verfahren, das auf der Basis von Begriffsdefinitionen (meist Phrasen), die von Experten in z.T. zeitraubender Arbeit – durchschnittlich vier Personenstunden pro Klasse – erstellt wurden, wobei (ohne dass dies im Detail offengelegt wurde) auch statistische Verfahren wie χ^2-Tests, schrittweise lineare Regression sowie ein proprietäres Gewichtungsverfahren, das mit positiven und

[28] http://www.lexisnexis.com [08.07.2004]

[29] vgl.: http://www.census.gov/epcd/www/sic.html [13.07.2004]

negativen Gewichten arbeitet, zur Anwendung kamen. Während dabei keinerlei Textnormalisierung (Stoppwortentfernung, automatische Depluralisation u.a. morphologische Verfahren) erfolgt, wird aber sehr wohl die Herkunft der Begriffsketten aus Überschrift, Abstract, Haupttext, Firmennamen-Feld usw. berücksichtigt. Ein Beispiel für eine solche Regeldefinition findet sich bei Quint (1999; vgl. *Tabelle 7-2*). Die ermittelten Gewichte werden auf einer 100-Punkte-Skala normiert, durch empirisch gewonnene Schwellenwerte bereinigt und drücken den Grad aus, zu dem ein Dokument "on-point" hinsichtlich eines bestimmten Deskriptors ist. Nach Wasson (2001) wurden in internen Tests Werte für Recall und Precision in der Grössenordnung von 90%–95% erzielt.

SCOPE NOTE: "JOINT VENTURES" targets joint partnerships between companies where a third, new company is formed and co-owned by the partners. The scope includes announcements of new ventures, failed ventures and joint venture banks.

VERY STRONG TERMS	STRONG TERMS	STRONG RELATED TERMS	WEAK TERMS	NEGATIVELY WEIGHTED TERMS
"concentrative" j/v	j-venture	form a venture	% of the arrangement	high school jv
"concentrative" j/vs	j-ventures	form the venture	% ownership	joint vision 2010
"concentrative" joint venture	j/venture	form ventures	% stake	junior varsity
"concentrative" joint ventures	j/ventures	formation of a venture	business combination	jv 2010
approve the j/v	joint venture	formation of the venture	business combinations	jv boys
approve the j/vs	joint venturer		agreement	jv girls
approve the joint venture	joint venturers		agreements	jv squad
to form joint venture	joint ventures		form joint	
to form joint ventures	joint-venture		form jointly	
	joint-venturer			
	joint-venturers			
	joint-ventures			

THRESHOLD = 4 {minimum score needed to tag a document}

HEADLINE COUNT = 2 {score given to a phrase matching in the headline segment}

BODY COUNT = 1 {score given to a phrase matching in the body segment}

Tabelle 7-2: *Lexis-Nexis* – Regel für den Begriff "joint ventures" (nach Quint 1999)

7.4.2 Northern Light

Von kaum einer der populären Suchmaschinen ist bekannt, dass sie ein Verfahren des automatischen Klassifizierens einsetzt. Eine Ausnahme stellt der – heute allerdings nicht mehr frei zugängliche – Suchdienst von *Northern Light* [30] (Cambridge, MA) dar, der es ermöglichte, Web-Ressourcen zusammen mit nicht am WWW verfügbaren Quel-

[30] http://www.northernlight.com [09.07.2004]

len (von Aggregatoren und Verlagen lizensierte elektronische Dokumente) in einer kombinierten Suche zu recherchieren (Attardi et al. 1999; Dumais et al. 2002; Krellenstein 2001; Northern Light 2003; Ward 1999). Klassifikatorische Elemente wurden dabei auf dreierlei Weise genutzt:

- Für das Relevance Ranking der Resultate einer Stichwortsuche (als eines von 17 Kriterien);
- Für die Darbietung der Suchresultate in einer hierarchisch gegliederten Ordnerstruktur ("custom research folders", vgl. *Abbildung 7-5*);
- Für die alternative Suchmöglichkeit nach Klassen bzw. das Browsing auf Basis der Klassenstruktur.

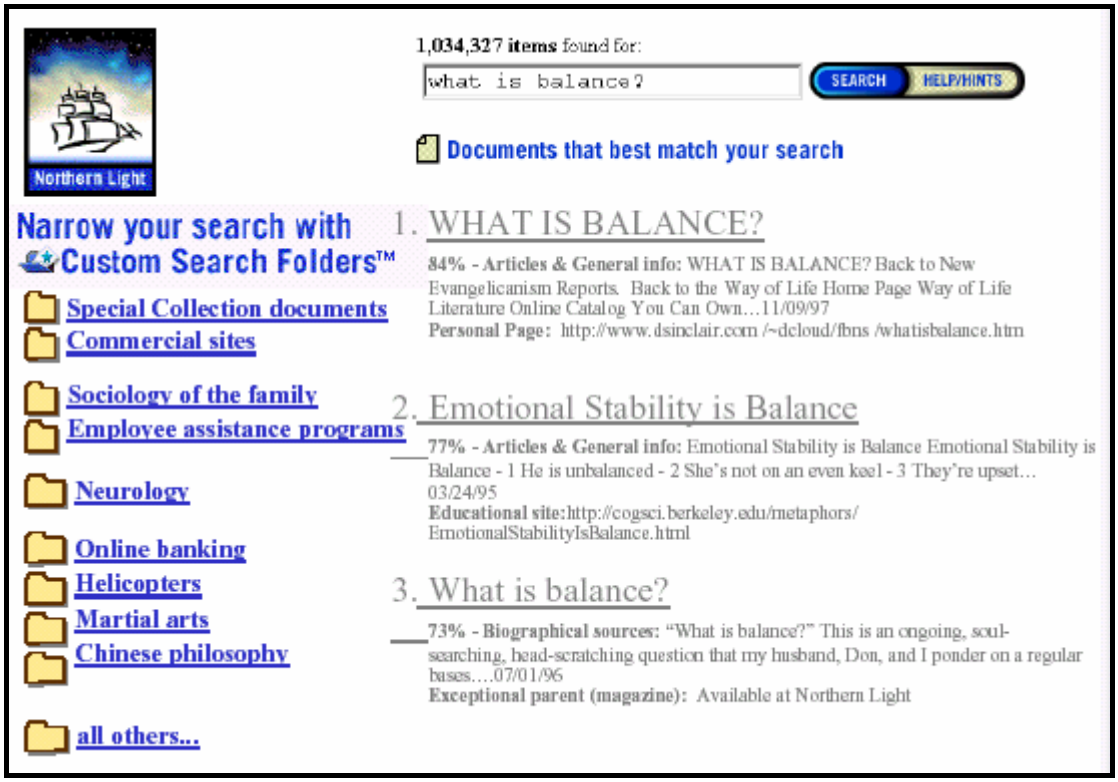

Abbildung 7-5: *Northern Light* – Beispiel für ein Suchresultat (Quelle: Ward 1999)

Dabei wurden mehrere Klassifikationssysteme verwendet. Das interessanteste davon ist zweifellos ein bei *Northern Light* selbst durch bibliothekarische Fachkräfte erstelltes und gewartetes thematisches System ("subject") mit 16 Hauptklassen und insgesamt rund 20.000 Sub-Klassen auf 7 bis 9 Hierarchieebenen. Weitere Schemata sind ein nur ca. 150 Klassen auf drei Hierarchiestufen umfassendes für Dokumententyp bzw. Genre ("type") sowie solche für "source" (z.B. Herkunft, Internetdomäne, übergeordnetes Werk/Zeitschrift), "language" und "region".

In methodischer Hinsicht gelangte ein proprietäres Verfahren zum Einsatz, das zwar nicht im Detail, doch zumindest in seinen Grundzügen offengelegt wurde. Danach handelte es sich eigentlich um die Kombination einer Reihe verschiedener Ansätze:

- Lineare Klassifikatoren, die durch überwachtes Lernen trainiert wurden;

- Regelbasierte Klassifikatoren, die manuell erstellt wurden und mit Techniken aus der künstlichen Intelligenz arbeiteten (offensichtlich die wichtigste der eingesetzten Komponenten);
- Metaregeln, die bewirkten, dass anstelle mehrerer Notationen aus demselben Zweig eine allgemeinere Notation zugeordnet wurde;
- Linguistische Techniken zur Spracherkennung;
- Manuelles und automatisches Mapping der Notationen einer Reihe existierender Systeme mit jenen der eigenen Schemata (insbesondere für bereits klassifikatorisch erschlossene Verlagspublikationen);
- Manuelles Klassifizieren bestimmter Web-Ressourcen.

Die Zahl der bei *Northern Light* klassifizierten Daten war beträchtlich. Ende der 1990er Jahre betrug dieses Volumen 141 Millionen Dokumente (davon 132 Millionen aus dem WWW), die überwiegend automatisch klassifiziert worden waren.

Für eine Menge von 20 Millionen Web-Ressourcen (95% automatisch, 5% manuell thematisch klassifiziert) wurde eine Precision von 90–95% angegeben, für 7 Millionen sonstiger elektronischer Dokumente (zu 100% durch Mapping oder automatische Verfahren klassifiziert) eine solche von 95%, ohne dass jedoch Details über die Art der Evaluierung bekanntgemacht wurden (Krellenstein 2001, sl.8). Aus anderer Quelle lässt sich jedoch erkennen, dass es sich dabei um Benutzerurteile handelte und dass nach den strengeren Anforderungen, die *Northern Light* selbst an das Verfahren stellte, wohl nur eine Precision von 60–65% erreicht wurde (Dumais et al. 2002).

Heute wird die mit einer ganzen Reihe weiterer Features versehene Suchmaschine inklusive des Klassifizierungsverfahrens vor allem als Intranetlösung für grosse Unternehmen vermarktet (Northern Light 2003).

7.4.3 Factiva

Der Informationsdienst *Factiva*[31] ist ein Tochterunternehmen von *Dow Jones* und *Reuters* und versteht sich als Anbieter von "world class global content" in Form von Datenbank- und Current-Awareness-Diensten einschliesslich deren Integration in betriebliche Informationsmanagement- und Workflowlösungen. 2002 verfügte dieser Dienst weltweit über 1,5 Millionen Subskribenten (hauptsächlich Firmen) und verarbeitete pro Monat 52 Millionen Artikel (Blumberg & Atre 2003, 6). Diese Texte stammen aus über 1.500 internationalen Zeitungen, 3.200 Zeitschriften und Fachmagazinen, den Newsfeeds von 500 Nachrichtenagenturen, nahezu 4.000 Websites (Unternehmens- und Wirtschaftsnachrichten) und einer Sammlung von über 30.000 Unternehmens- und 340.000 Personenprofilen (Factiva 2004).

Seit 1999/2000 wird für die Erschliessung aller Texte (in 22 Sprachen) das auf Basis der langjährigen Erfahrungen von Dow Jones und Reuters und z.T. in Anlehnung

[31] http://www.factiva.com [08.07.2004]

an das *North American Industry Classification System* (NAICS)[32] erarbeitete proprietäre System *Factiva Intelligent Indexing* verwendet. Dabei handelt es sich um ein teilweise hierarchisch strukturiertes System ("coding system", "taxonomy") mit folgenden Komponenten (Sykes 2001; 2003):

- 760 "industry codes" zur Identifizierung von Wirtschaftszweigen (5 Hierarchiestufen);
- 470 "news subject codes" zur Codierung von Nachrichtenthemen (4 Hierarchiestufen);
- 370 "region terms" zur Kennzeichnung von Ländern, Bundesstaaten/Provinzen (USA bzw. Kanada) und übernationalen Einheiten (z.B. "südliches Afrika");
- 300.000 "company codes" zur eindeutigen Kennzeichnung von Unternehmungen.

Bei dem von *Factiva* eingesetzten Klassifizierungsverfahren handelt es sich, ähnlich wie im Fall von *Northern Light*, um einen methodischen Mix:

- regelbasierter Klassifikator zur Identifizierung schwieriger Firmennamen;
- auf der Basis von Trainingsdokumenten lernender *k*-NN Klassifikator für die Kategorisierung nach Themen, Regionen, Wirtschaftszweigen (mindestens 30 Trainingsdokumente pro Klasse); dabei werden die mit hoher Sicherheit zugeordneten Notationen automatisch übernommen und die unsicheren Kandidaten-Notationen manuell überprüft;
- Mapping-Techniken für Dokumente, die bereits nach anderen Systemen erschlossen sind (z.B. Reuters- und Dow Jones-Ressourcen);
- natürlichsprachliches Textanalyseverfahren zur Extrahierung von Wörtern/Phrasen, die als Kandidaten für Firmennamen infrage kommen und mit einem Verzeichnis von Namensvarianten von Firmen verglichen werden; wenn dabei keine Übereinstimmung erzielt wird, so erfolgt eine manuelle Prüfung (Erstellung eines neuen Eintrags oder einer neuen Namensvariante).

Ein wesentlicher Unterschied besteht darin, dass im Fall von *Factiva* das Verfahren zum Grossteil nicht selbst entwickelt wurde, sondern vorwiegend Produkte eines kommerziellen Anbieters – *Inxight*[33] – zur Anwendung kommen (*Thing Finder; Categorizer*). Diese Software nimmt auch für sich in Anspruch, eine grosse Zahl von *Sprachen* zu unterstützen (Inxight 2002; 2004).

Eine 2002 durchgeführte Evaluierung mit Hilfe von Fokusgruppen erfahrener und gelegentlicher Benutzer erbrachte für Precision und Recall Werte von 86% bzw. 85% (keine näheren Details bekannt); da für die Kunden Precision besonders wichtig ist, wird die Erhöhung der Precision auf 90% angestrebt (Sykes 2003, 6).

Aus der Firmen-Homepage geht hervor, dass die genannten halbautomatischen Klassifizierungskomponenten auch im Wege verschiedener Dienstleistungsmodelle angeboten werden (inklusive der proprietären Taxonomie bzw. der Entwicklung massgeschneiderter Systeme für Kunden).

[32] http://www.census.gov/epcd/www/naics.html [13.07.2004]

[33] *Inxight* ist ein "spun-off" von Xerox PARC; http://www.inxight.com [06.05.2004]

7.4.4 INFOMINE

Eines der neuesten relevanten Projekte ist die automatische Klassifizierung von *INFOMINE*,[34] einer an der University of California (Riverside) betriebenen virtuellen Bibliothek wissenschaftlicher Web-Ressourcen (Frank & Paynter 2004). Diese über 20.000 Dokumente waren zwar von bibliothekarischen Fachkräften mittels LCSHs sachlich erschlossen worden, doch stellte sich ab Erreichen einer gewissen Grösse das Fehlen einer klassifikatorischen Erschliessungskomponente als gravierender Mangel für die Präsentation bzw. Suche am Web heraus. Daher wurde ein Projekt zur automatischen Zuordnung von Notationen der *LCC Outline* (einer LCC-Version mit den ca. 4.200 obersten Klassen dieses Systems) zu den Dokumenten initiiert. Als Grundlage dafür sollte – nicht zuletzt aufgrund der Ergebnisse von Larson (1992; vgl. Abschnitt 7.1.1) nur die bereits vorhandene LCSH-Erschliessung (und nicht z.B. der Titel oder weitere Elemente aus den Dokumenten) dienen.

Aufgrund der mit *Support Vector Maschinen* (vgl. Abschnitte 2.5.2 und 2.6.3) erzielten Erfolge wurde ein solches Verfahren gewählt und dazu *Open Source*-Software eingesetzt. Da SVMs binäre Klassifikatoren sind, hier aber ein Fall von *multi-label classification* (vgl. 2.1.2) vorlag, wurde paarweise vorgegangen. Bei der paarweisen Klassifizierung wird ein binärer Klassifikator trainiert, zwischen jedem Paar von Klassen zu unterscheiden, wobei für das Training nur Dokumente aus diesen Klassen verwendet werden. Das bedeutet, dass eine grosse Zahl von Klassifikatoren erstellt werden muss.[35] Um die hierarchische Struktur der LCC in den Prozess miteinzubeziehen, wurde jedes Dokument zunächst mit einem "root node classifier" einer der 21 Hauptklassen zugeordnet. In dieser teilte ein "internal node classifier" das Dokument einer Unterklasse zu; dieser Prozess wurde solange wiederholt, bis keine weitere Untergruppe mehr verfügbar war ("leaf node") oder eine weitere Unterteilung ein schlechteres Ergebnis erbracht hätte. Um auch zu einer ranggeordneten Liste der zu einem Dokument am besten passenden Notationen zu gelangen, wurden Konfidenzwerte für die anderen möglichen Pfade durch die Hierarchie berechnet und als Relation zu "1", dem Wert für den besten Pfad (d.h. die beste Notation), ausgedrückt.

Als Trainingsdokumente wurden ca. 870.000, über die rund 4.200 Klassen der *LCC Outline* verteilte bibliographische Datensätze aus dem Katalog der UB in Riverside verwendet. Im Zuge der Datennormalisierung wurden die vorhandenen LCC-Notationen auf die jeweils spezifischste übereinstimmende Notation aus der *Outline* abgebildet; die LCSHs wurden in Kleinbuchstaben umgewandelt und gekürzt (Entfernung von Text in Klammern sowie aller "subdivisions").

---

[34] http://infomine.ucr.edu/ [15.07.2004]

[35] Bei *m* Klassen sind dies *m * (m – 1) / 2* Klassifikatoren.

Aus der Trainingsmenge wurden 50.000 Datensätze zufällig gezogen und als Validierungsmenge reserviert; mit dem Rest erfolgte ein mehrstufiges Training der Klassifikatoren.[36] Die Ergebnisse der nachfolgenden Validierungstests zeigten, dass bei Verwendung von 800.000 Trainingsdokumenten 80,27% der Dokumente auf der obersten Hierarchiestufe korrekt zugeordnet wurden; auf der untersten (siebenten) Hierarchiestufe waren dies nur mehr 16,12% (Fehler, die auf höheren Hierarchiestufen passieren, können in unteren nicht mehr korrigiert werden). Bei Mehrfachklassifizierung wurden 79,2% der "wahren" Notationen unter den erstgereihten 10 und 82,2% unter den erstgereihten 15 der automatisch zugeteilten registriert.

Schliesslich wurden die rund 20.000 Dokumente aus *INFOMINE* mit diesem Verfahren automatisch klassifiziert, wobei jedoch pro Dokument nur *eine* Notation vergeben wurde. Die zugeordneten Notationen werden nunmehr in einem Interface für das thematische Browsing in dieser virtuellen Bibliothek eingesetzt. Als Probleme bei der Anwendung des Klassifizierungsverfahrens auf die neu zu klassifizierenden Dokumente wurden bisher folgende Fälle erkannt:

- LCSHs, die in den Trainingsdokumenten nicht aufgetreten waren (4% der Kollektion);
- LCSHs, die in den Trainingsdaten nur einmal auftraten und selbst dann nur von sekundärer Bedeutung für das betreffende Trainingsdokument waren;
- Sehr allgemeine LCSHs wie "History" und "United States", bei denen – wie auch bei den übrigen – im Prozess der Attributauswahl die "subdivisions" entfernt wurden, wodurch der thematische Kontext verloren ging.

7.4.5 Sonstige Anwendungen und Projekte

Alexandria / Pharos. Von einer im Rahmen des Projekts *Alexandria Digital Library*[37] erstellten Prototyps namens *Pharos* berichteten Dolin et al. (1988). Dieses System war für die Verarbeitung bzw. das Retrieval einer grossen Zahl inhaltlich und formal sehr heterogener geographischer elektronischer Informationsressourcen vorgesehen. Es sollte einerseits deren automatische Klassifizierung anhand mehrerer vorgegebener hierarchischer Systeme durchführen und andererseits eine klassifikationsbasierte Abfragekomponente anbieten. Der in der genannten Arbeit beschriebene Prototyp basierte auf der Verwendung

- der *LCC Outline* (einer kürzeren LCC-Version; vgl. Abschnitt 7.4.4),
- von 1,5 Millionen MARC-Datensätzen aus dem Bibliothekskatalog der University of California (Santa Barbara) als Trainingsmenge,
- von 1.500 Usenet Newsgruppen als zu klassifizierendem Korpus (wobei jeweils die gesamte Newsgruppe als "Dokument" betrachtet wurde),

[36] Bei 800.000 Datensätzen wurden hiefür 19 Tage benötigt (Frank & Paynter 2004, 223).

[37] http://www.alexandria.ucsb.edu/ [13.07.2004]

- einer Attributextraktion mittels "Latent Semantic Indexing" (vgl. Abschnitt 2.4.2),
- und eines linearen Batch-Klassifikators.

Die Benutzerschnittstelle sah kein Browsing in der Klassenstruktur, sondern eine automatische Zuordnung der eingegebenen Stichwörter zu den Klassen der LCC (mittels einer probabilistischen Retrievalkomponente) und die ranggeordnete Ausgabe der diesen Klassen am besten entsprechenden Dokumente vor. Zwar stellte das Entwicklerteam eine öffentlich zugängliche Webseite zum Ausprobieren dieses Systems zur Verfügung, doch wurde gleichzeitig angeregt, dabei möglichst spezifische Begriffe zu verwenden, was darauf schliessen lässt, dass bei weniger spezifischen Termen Probleme mit der Disambiguierung aufgetreten sein dürften.

Die Evaluierung dieses Prototyps konnte offensichtlich nur durch eine intellektuelle Beurteilung der zur jeweiligen Frage gefundenen Newsgruppen erfolgen, nicht aber durch einen Test auf Precision und Recall des Klassifikators, da die klassifizierten Ressourcen nicht zuvor intellektuell mittels der LCC erschlossen worden waren.

COMPCAT. Hiebei handelt es sich um ein 2002 begonnenes Projekt des Instituts für Informationsverarbeitung des *Consiglio Nazionale delle Ricerche*, Pisa (Sebastiani 2001; 2002a). Ziel war die Erstellung eines interaktiven Klassifikators für die Erschliessung von wissenschaftlichen Artikeln im Rahmen des Aufbaus einer digitalen Bibliothek für das Fachgebiet Informatik. Das System sollte den Anwendern pro Artikel eine ranggeordnete Liste von Notationsvorschlägen aus dem Klassifikationsschema der *ACM Computing Reviews* präsentieren. Als Trainings- und Testdokumente waren die nach diesem System erschlossenen Jahrgänge der *ACM Digital Library*[38] (seit 1998, da die in diesem Jahr neu herausgegebene Version des Schemas verwendet wurde) vorgesehen. In methodischer Hinsicht waren Ansätze zur Nutzung der durch die hierarchische Struktur ausgedrückten Information sowie die Kombination mit einem clusteranalytischen Verfahren geplant.

Cora. Unter der Bezeichnung *Cora* war mehrere Jahre eine fachspezifische Suchmaschine für Forschungsarbeiten aus dem Gebiet Computerwissenschaft öffentlich zugänglich (McCallum et al. 1999; 2000). Dabei wurde ein fachspezifisches Harvesting mittels eines lernenden induktiven Prozesses ("topic spidering", "reinforcement learning") eingesetzt, um über 50.000 einschlägige Ressourcen aus dem WWW, v.a. von den Webseiten der führenden (amerikanischen) Informatik-Institute, zu sammeln bzw. dabei nichtwissenschaftliches Material (z.B. Lehrpläne, Verwaltung) auszuscheiden. Primär wurden dabei Arbeiten im Postscript-Format gesammelt, in "plain text" verwandelt und dann weiterverarbeitet. Eine automatische Klassifizierungskomponente, die einen Nai-

[38] http://www.acm.org/dl/ [13.07.2004]

ven Bayes-Algorithmus verwendete, wurde eingesetzt, um die Dokumente nach einem selbst erstellten, relativen groben, aber doch mehrstufigen hierarchischen Schema zu klassifizieren. Da keine bereits klassifizierten Trainingsdokumente zur Verfügung standen, wurde der Klassifikator zunächst mittels manuell gebildeter Klassendefinitionen trainiert und im Laufe des Klassifizierungsprozesses iterativ mit neuen Attributen gespeist ("bootstrapping iterations"), was im besten Fall zu einer Genauigkeit von bis zu 66% führte. Das zunächst von *Just Research* gemeinsam mit der Carnegie Mellon Universität (beide Pittsburgh, PA) durchgeführte Projekt wurde später von *Whizbang! Labs* übernommen (Gietz 2001), einem Unternehmen, das 2002 von der bereits mehrmals erwähnten Firma *Inxight* gekauft wurde. Trotz mehrerer Versuche des Autors war die Suchmaschine von *Cora* nicht mehr am WWW auffindbar.

eBay. Das seit 1995 bestehende WWW-Auktionsportal *eBay*[39] nutzt ein selbst erstelltes Klassifikationssystem für die Präsentation seiner Bestände und als Suchhilfe für die Benutzer. Oblag es ursprünglich ausschliesslich dem jeweiligen Anbieter einer Auktionsware, diese in ein hierarchisches Gerüst von 2.900 Klassen einzuordnen (Kwasnik & Liu 2000, 373), so wird heute dafür eine automatische Kategorisierung mithilfe der kommerziellen Software *Texis Categorizer* des Herstellers *Thunderstone*[40] vorgenommen (Thunderstone o.J.) Dabei handelt es sich um ein regelbasiertes, lernendes Verfahren, das in der Lage ist, auch Anmerkungen und Kategorienvorschläge der Anbieter zu berücksichtigen bzw. in den weiteren Lernprozess (der mit etwa 20 Beispieldokumenten pro Klasse beginnt) aufzunehmen (Lamont 2003). Der Input neuer Auktionsware beläuft sich heute auf etwa 2 Millionen Artikel pro Tag; der Umfang des Klassifikationssystems wird inzwischen mit rund 45.000 Klassen angegeben (Knox 2003).[41] Es ist ein von Firmenmitarbeitern pragmatisch erstelltes und laufend nach Marktbedürfnissen erweitertes hierarchisches Schema,[42] das allerdings viele "klassische" Kriterien eines Klassifikationssystems (wie z.B. Homogenität der Klassen, wechselseitiges Ausschliessen, konsistente Struktur) ignoriert bzw. verletzt (Kwasnik & Liu 2000; Schultz o.J.)

GRACE. Dieses Akronym steht für *Grid Search And Categorization Engine* und bezeichnet ein von 2002 bis 2005 laufendes Projekt[43] aus dem 5. Rahmenprogramm der EU-Initiative *Information Society Technologies* (Haya et al. 2003; Scholze 2003).

[39] http://www.ebay.com, http://www.ebay.de, http://www.ebay.at (u.a.) [23.03.2004]

[40] http://www.thunderstone.com [09.07.2004]

[41] Dieser Zahl ist möglicherweise mit Vorsicht zu begegnen, da auf der in der folgenden Fussnote zitierten Webseite von *eBay* nichts auf diese Grössenordnung hinweist.

[42] http://listings.ebay.com/aw/listings/overview.html [13.07.2004]

[43] http://www.grace-ist.org [13.07.2004]. Die Projektpartner sind Telecom Italia Lab, CERN, die israelische Firma Virtual Self, die Universität Sheffield Hallam sowie die Universitätsbibliotheken Stockholm und Stuttgart.

GRACE zielt auf die Erstellung einer Suchmaschine für einen wissenschaftlichen Nutzerkreis unter Verwendung des Konzepts des *Grid-Computing* ab, das im wesentlichen ein sehr weit verteiltes Daten-, Rechner- und Applikationsnetzwerk beschreibt. Das Programm von *GRACE* sieht hinsichtlich einer Kategorisierung der Ressourcen in erster Linie die Verwendung clusterbildender statistischer Verfahren und die Generierung der Klassenbezeichnungen mittels linguistischer Techniken in den Sprachen Deutsch, Englisch, Schwedisch, Italienisch) vor. Daneben soll aber auch eine Kategorienbildung anhand vorgegebener Klassifikationssysteme bzw. Thesauren realisiert werden, wobei z.B. an den *High Energy Physics Thesaurus* gedacht ist.

Snowfox. Dies ist die Bezeichnung für einen experimentellen Web-Katalog,[44] der auf einer "Ontologie" (gemeint ist wohl nur ein hierarchisches Kategorienschema) von 7.500 Themen beruht und mit dem Verfahren *Snowfox Relevance Engine* erstellt wurde. Dieses nimmt für sich in Anspruch, Trainingsdokumente "vollautomatisch" (d.h. auf der Basis nur einiger Anfangsbegriffe) aus dem WWW filtern und diese mittels eines als "haystack algorithm" bezeichneten Verfahrens so effizient von störenden Textelementen (Rauschen) befreien zu können, dass schliesslich nur mehr das für einen lernenden Klassifikator relevante Vokabular übrig bleibt. Die zugrundeliegende Technik basiert möglicherweise auf einem Verfahren der Mustererkennung, wird aber aus "patentrechtlichen" Gründen nicht näher ausgeführt. Sowohl die Homepage als auch das "white paper" zu diesem Projekt (Gower 2002) legen aufgrund der dort getätigten Behauptungen sowie auffallend vieler Schreib- und Flüchtigkeitsfehler gewisse Zweifel an der Seriosität des Ansatzes nahe.

Varia. Abschliessend seien noch kurz einige Ansätze und Projekte erwähnt, mit denen im Rahmen dieser Arbeit keine detailliertere Auseinandersetzung möglich war bzw. für die keine ausreichende Informationsgrundlage ermittelt werden konnte, die weiter zu verfolgen sich aber vielleicht lohnen könnte:

- ***Begriffsanalyse:*** Neuss & Kent (1995) schlugen ein System zur Erfassung und Erschliessung von Web-Ressourcen vor, das auf einer Interpretation dieser Ressourcen nicht als Objekte, sondern als Begriffsklassen ("conceptual classes") beruht und mit Begriffsskalen ("conceptual scales") operiert, die den Facetten synthetischer Klassifikationssysteme (wie etwa der CC) entsprechen.
- ***DDC / Wirtschaftswissenschaften.*** Im Rahmen einer vor kurzem in Korea durchgeführten experimentellen Studie kehrten Chung & Noh (2003) zum Ansatz eines manuell erstellten "Lexikons" der Klassenbenennungen zurück. Dabei wurden rund 6.700 englischsprachige Web-Ressourcen mit einem linearen Klassifikator nach 757 Klassen der DDC (aus "330" = Economics) klassifiziert, wobei eine Precision von 77% erzielt werden konnte. Durch Verwendung eines Teils der so klassifizierten Dokumente als Trainingskollektion konnte bei Anwendung eines *k*-NN Verfahrens auf die übrigen eine Steigerung auf 96% Precision er-

[44] http://www.snowfox.com/ [23.03.2004]

zielt werden. Dieser hohe Wert ist m.E. jedoch als Artefakt aus der gewählten Versuchsanordnung zu interpretieren.

- ***DR-LINK:*** Das im Rahmen dieses Projekts der *School of Information Studies*, Syracuse University, eingesetzte Verfahren verwendet zwar zur Erstellung der Klassen einen Cluster-Algorithmus, setzt aber als ersten Schritt einen interessanten Ansatz ein, bei dem zunächst die Wörter in den Dokumenten mit (multiplen) "subject codes" aus dem *Longman's Dictionary of Contemporary English* – einem Schema aus 124 Haupt- und 250 Unterklassen – verknüpft werden. Mittels einer Software zur natürlichen Sprachverarbeitung werden sukzessive die Mehrdeutigkeiten reduziert und durch Einbeziehung des Kontexts die "richtigen" Codes ermittelt. Der so mit normiertem Vokabular ausgestattete Text wird sodann geclustert (Liddy et al. 1994).
- ***OASIS:*** Dieses EU-Projekt zielte zwar darauf ab, einen Web-Katalog mittels eines clusteranalytischen Verfahrens zu erstellen, doch wurde in einem Teilprojekt auch ein klassifikationsbasierter Ansatz zum thematisches Vorfiltern beim Web-Harvesting verfolgt (Nekrestyanov et al. 1999).
- ***PEKING:*** In diesem 2001–2002 an den Universitäten Nijmegen (Niederlande) und Barcelona (Spanien) durchgeführten EU-Projekt[45] ging es einerseits um die Ablösung der manuellen Klassifizierung steuerrechtlicher Dokumente durch ein automatisiertes System, die Überführung des bereits klassifizierten Korpus in eine neue Version des Klassifikationsschemas (Dumais et al. 2002) und v.a. auch um die Probleme der *mehrsprachigen* Klassifizierung ("cross-lingual classification", "bi-lingual classification") mit einer Reihe von Experimenten mit Dokumenten in englischer und spanischer Sprache (Bel et al. 2003).
- ***Portugal / Rechtsinformationssystem:*** Ein System zur automatischen Kategorisierung von juristischen Web-Dokumenten (nach einem fachspezifischen Klassifikationssystem), das mit einem Klassifikator auf der Basis künstlicher neuronaler Netze arbeitet, wurde an der Universität Evora (Portugal) erstellt; dabei wurde eine Genauigkeit der Zuordnung von 90% erzielt (Quaresma & Rodrigues 2002).
- ***Theseus / Teseo:*** Dieses an der Universität Pisa durchgeführte Projekt zur Kategorisierung von Web-Ressourcen basiert auf dem "categorization by context"-Ansatz, d.h. der Einbeziehung der Hyperlinks in den zu klassifizierenden Ressourcen bzw. der durch diese Links assoziierten Dokumente (Attardi et al. 1999).
- ***Yahoo! / LookSmart:*** Eine Reihe von Arbeiten beschäftigte sich mit der automatischen Klassifizierung von Web-Dokumenten, Pressemeldungen etc. nach den umfangreichen hierarchischen Schemata der bekannten Web-Kataloge *Yahoo!*[46] (Ceci & Malerba 2003; Chekuri et al. 1997; Mladenic 1998; Mladenic & Grobelnik 1999) und *LookSmart*[47] (Chen & Dumais 2000; Dumais & Chen 2000). Bei diesen Studien ging es jedoch primär um methodische Fragen und nicht um die Erarbeitung einer praktischen Anwendung für den Gebrauch am WWW.

[45] http://www.cs.kun.nl/peking/ [25.03.2004]

[46] http://www.yahoo.com/ [16.05.2004]

[47] http://search.looksmart.com/ [16.05.2004]

8 Diskussion und Ausblick

In diesem Kapitel werden die bisher dargestellten Verfahren und Anwendungen kurz diskutiert und einige weitere im Zusammenhang mit dem automatischen Klassifizieren relevante Aspekte angesprochen.

8.1 Zur Methodik des automatischen Klassifizierens

Der methodische Abriss in Kapitel 2 hat gezeigt, dass zum gegenwärtigen Zeitpunkt kein Zweifel an der Akzeptanz des Paradigmas vom automatischen Klassifizieren als der *Vorhersage einer Klassenzuordnung mittels eines auf der Basis des maschinellen Lernens induktiv erstellten Klassifikators* besteht. In der Tat stellt Fabrizio Sebastiani, eine Autorität auf dem Gebiet der automatischen Textklassifizierung, dazu apodiktisch fest, "The effectiveness of automatically built classifiers now rivals that of human classifiers" (2002a, sl.10). Dies mag richtig sein – insbesondere, wenn man die bekanntermassen nicht perfekte Leistung menschlicher Indexierer ins Kalkül zieht –, doch steht der "Beweis" für diese Behauptung noch in zweierlei Hinsicht aus. Zum einen gibt es noch kaum Anwendungen, die für sich in Anspruch nehmen können, wirklich *vollautomatisch* abzulaufen; zum anderen fehlen erfolgreiche praktische Implementierungen mit *sehr grossen Datenmengen* und *sehr komplexen Klassifikationssystemen*. Marcia Bates, eine Autorität auf dem Gebiet der Informationswissenschaft, warnt allerdings vor allzu grossem Optimismus hinsichtlich einer Vollautomatisierung der inhaltlichen Erschliessung (1998, 1186):

> ... it is not difficult to get that first 70% in retrieval systems – especially with small prototype systems. The last 30%, however, is infinitely more difficult. [...] Information retrieval also poses serious scalability problems; small prototype systems are often *not* like their larger cousins. Further, user needs vary not just from one time to another, but from one subject domain to another. Optimal indexing and retrieval mechanisms may vary substantially from field to field.

Vor diesem Hintergrund darf man vor allem der Realisierung des Systems *GERHARD II* mit Spannung entgegensehen (vgl. auch Abschnitt 8.4).

Welcher methodische Ansatz bei einem allfällig geplanten Erschliessungsprojekt (z.B. einer Bibliothek bzw. Informationseinrichtung) zu wählen wäre, ist wohl schwer nur auf der Basis der rezipierten Literatur und ohne grundlegende Informatik-Kenntnisse entscheidbar. Einen Hinweis mag aber z.B. die aktuelle Studie von Frank & Paynter (2004) geben, in der, wie in Abschnitt 7.4.4 dargestellt wurde, ein *SVM-Verfahren* für die hierarchische Klassifizierung mit der LCC und mit bibliothekarischen Datensätzen als Trainingsmenge gearbeitet wurde. In der Praxis wird diese Frage sicherlich auch davon abhängen, ob man sich für kommerzielle Software entscheidet – in diesem Fall

ist die methodische Transparenz wahrscheinlich begrenzt – oder es auf sich nimmt, selbst ein (frei erhältliches) Softwarewerkzeug zu suchen, einzurichten und zum Laufen zu bringen. Die korrekte Vorgangsweise, was Trainings- und Testdokumente bzw. die Kriterien für die Güte des eingesetzten Verfahrens betrifft, sollte dabei jedenfalls ausser Frage stehen.

8.2 Die Projekte an der Universität Lund

Nordic WAIS/WWW. Das Projekt kann als einer der ersten Versuche gelten, ein automatisiertes Klassifikationsverfahren für die Verbesserung des Zugriffs auf elektronische Dokumente einzusetzen (Srishaila 2001, 5–6). Weitere NetLab-Projekte gehen darauf zurück: "Many of the themes we have worked with later can be seen in embryonic form there" (Ardö et al. 2002, 2). In seiner frühen Kompilierung einschlägiger Unternehmungen kategorisiert McKiernan *W4* immerhin als "large scale project" (1996, 28).

Aus heutiger Sicht ist *Nordic WAIS / WWW* vor allem von historischer Bedeutung. Das eingesetzte Klassifikationsverfahren selbst ist äusserst einfach (simple Wort-Übereinstimmung mit Gewichtung). Überdies wird trotz der zahlreichen zu diesem Projekt vorliegenden Publikationen nicht klar, wie das Verfahren im Detail funktionierte, da zumindest folgende Punkte nicht offengelegt wurden:

- die konkrete Herkunft des als "UDC vocabulary" bezeichneten Wortmaterials (handelte es sich um Stichwörter bzw. Phrasen aus den Klassenbenennungen oder auch um Wörter bzw. Phrasen aus den entsprechenden Einträgen des Sachregisters?);
- die konkrete Vergabe der Gewichte (welche Gewichte wurden für welche "Gruppe" vergeben?);
- die Funktionsweise des heuristischen Verfahrens, mit dem die Auswahl der tatsächlich zugeteilten Notationen erfolgte (z.B. Bestimmung der Schwellenwerte).

DESIRE II. Das im Rahmen von *DESIRE II* eingesetzte Klassifizierungsverfahren ist abermals nur ein einfacher Matching-Algorithmus, d.h. ein regelbasierter Ansatz, der als methodisch eher unbefriedigend zu bewerten ist. Dies war wohl auch die Meinung einer amerikanischen Gutachterin, die höflich meinte, "there are other approaches to automatic classification and categorization that may be considered", aber dennoch unmissverständlich auf modernere Verfahren verwies (Srinivasan o.J.) Auch der zweite Gutachter erwähnte mögliche methodische Alternativen, assoziierte aber dabei eher ein hyperlinkbasiertes Clustering (Jansson o.J.) Die eher intuitiv wirkende Vorgangsweise bei der Gewichtung (Srinivasan o.J.) – vom Projektteam nur vage mit früheren Erfahrungen und "Heuristik" beantwortet (ibid.) – lässt vermuten, dass einfach so lange probiert wurde, bis eine zufriedenstellende Lösung gefunden war. Überdies zeigten auch die einge-

setzten Evaluierungsmethoden manche Schwächen (Übereinstimmung mit intellektueller Klassifizierung; Expertenurteile).

Trotz dieser methodischen Unzulänglichkeiten handelte es sich bei dieser Erprobung des automatischen Klassifizierens – auch nach Meinung der zitierten Gutachter – ohne Zweifel um ein signifikantes Projekt, das zeigte, welches Potential in der Anwendung eines etablierten Klassifikationssystems für die automatische Erschliessung von Web-Dokumenten liegen kann. Insbesondere die demonstrierte Anreicherung der Klassenbenennungen durch einen Thesaurus ist m.E. ein interessanter methodischer Einfall, selbst wenn im konkreten Fall dabei die Homonymkontrolle nicht ausreichend gelungen ist (vgl. Evaluierung). Auch die ausführliche, obzwar nicht leicht zu rezipierende Dokumentation der Vorgangsweise verdient positive Erwähnung.

Bedauerlich ist allerdings, dass bereits wenige Jahre nach der Durchführung dieses Projekts ein Teil der zugehörigen Webseiten nicht mehr erreichbar ist – insbesondere die vermutlich lehrreichen Varianten der aufgrund unterschiedlicher Parameter resultierenden Browsingstrukturen sowie der interaktive WWW-Klassifikator (beides Ardö & Koch 2000).

Engine-e. Dieses Projekt ist wohl vor allem primär als Nachnutzung des in *DESIRE II* entwickelten Konzepts zu sehen. Für das automatische Klassifizieren wurde wiederum auf das bereits früher verwendete simple Matching-Verfahren gesetzt. Der bisher vorliegende Prototyp bedarf nach Ansicht der Entwickler noch weiterer Verbesserungen (Begriffsliste, Zuordnungsmechansimus). Ob diese noch realisiert werden, ist nach aktuellen Informationen auf der Hauptseite von *Engine-e*,[1] wonach derzeit keine Aktualisierung bzw. Entwicklung stattfinde, zumindest fraglich.

8.3 Wolverhampton Web Library

WWLib-TOS und "Old ACE". *WWLib-TOS* bzw. *Old ACE* stellten einen der ersten Versuche zur automatischen klassifikatorischen Erschliessung von Webseiten in grösserem Umfang und unter Verwendung einer bibliothekarischen Universalklassifikation (DDC) dar. Im methodischen Ansatz des relativ einfachen, regelbasierten Verfahrens findet sich zwar manche grundsätzliche Ähnlichkeit wie bei *DESIRE,* doch weicht die Vorgangsweise in vielerlei Hinsicht von jener des überdies fachlich orientierten EU-Projektes ab. Was das die Erstellung eines die Klassen repräsentierenden Vokabulars betrifft, so war man bei *DESIRE* durch die Verfügbarkeit eines klassifizierten Thesaurus, der eine Anreicherung der Klassenbenennungen ermöglichte, sicherlich im Vorteil gegenüber dem etwas zu einfach und pragmatisch zusammengestellten DDC-Vokabular

[1] http://engine-e.lub.lu.se [09.06.2004]

von *WWLib* – ein möglicherweise von den Entwicklern zunächst auch unterschätzter Aspekt. Im Hinblick auf das eingesetzte Matching-Verfahren und die zahlreichen optionalen Gewichtungsmöglichkeiten wirkt *Old ACE* ein wenig "more sophisticated" als die für *AE* entwickelte Lösung. Die leider nicht sehr gut dokumentierten Evaluierungstests weisen jedoch auf eine nicht wirklich befriedigende Klassifizierungsleistung hin, die vermutlich nur durch das Ausscheiden einer grösseren Zahl von Web-Ressourcen als nicht-klassifizierbar "verbessert" werden konnte.

WWLib-TNG und ACE. Bemerkenswert an diesem – bisher augenscheinlich (noch) nicht in die Realität eines am Web benutzbaren Dienstes umgesetzten – Ansatz sind zum einen die präzise und programmiernahe Modellierung und zum anderen die Intention, durch Berücksichtigung der hierarchischen Klassifikationsstruktur eine Art Filtereffekt zu erzielen und so mehrdeutige Begriffe quasi "kontext-sensitiv" disambiguieren zu können. Die dafür gewählte Vorgangsweise impliziert allerdings, dass nur solche Dokumente dem Matching-Verfahren für eine spezifischere Klasse unterzogen werden, die zuvor die Prüfung auf Übereinstimmung mit dem Vokabular der jeweils übergeordneten, allgemeineren Klasse "bestanden" haben. Dies wird etwa am Beispiel des Polysems "litter" – Strassenabfälle, Streu <Stall>, Sänfte/Trage, Wurf <Tiere> – in der fiktiven Klasse "cat" mit dem Argument begründet, dass eine Seite über Abfälle bei der Matching-Prüfung gegen das Vokabular der übergeordneten Klasse "animal" nur unwahrscheinlich einen signifikanten Wert erzielen und somit gar nicht zur der Matching-Prüfung gegen das Vokabular der Klasse "cat" gelangen würde (Jenkins et al. 1997, Kpt. 4). Ein Versuch, diese doch etwas gewagt anmutende Hypothese zu falsifizieren, wurde m.W. freilich erst gar nicht unternommen.

Auch die bisher vorliegenden Evaluierungsergebnisse sind nicht umfangreich bzw. präzise genug, um daraus Schlüsse auf die tatsächliche Leistungsfähigkeit des Verfahrens ziehen zu können.

Weitere Experimente mit ACE. Die neueren im Zusammenhang mit ACE vorgenommenen Experimente weisen v.a. in methodischer Hinsicht auf "neue Wege" (maschinelles Lernen, Verwendung domänenspezifischer Ontologien). Welche praktische Relevanz (Umsetzung in ein Anwendersystem) diesen Versuchen zukommt, kann zur Zeit allerdings nicht abgeschätzt werden.

8.4 GERHARD

Das DFG-Projekt GERHARD. Koch (1998b, 334) beurteilte *GERHARD* wie folgt:

> Es handelt sich um den umfassendsten und am tiefsten erschlossenen Katalog, der aufgrund der automatischen Erstellung ausserdem sehr schnell sein kann. Grösser, wesentlich aktuel-

> ler, mehr systematisch erschlossen und mit erheblich weniger Aufwand geschaffen als zum Beispiel Yahoo!

Aus heutiger Sicht zeigt das Zitat vor allem das *Potential*, das *GERHARD* bei einem Dauerbetrieb gehabt hätte. Dazu war ursprünglich auch beabsichtigt gewesen, Sponsoren und Bannerwerbung für die Dauerfinanzierung zu gewinnen, mit Verlagen zu kooperieren u.a.m. (Wätjen et al. 1998, 32–33). Offenbar wurde aber schon sehr bald dem Betreiben des Nachfolgeprojektes der Vorzug eingeräumt. Dieses ist jedoch bis heute noch nicht fertiggestellt, sodass die bestehende Webseite von *GERHARD*, die auf dem Stand von 1998 "eingefroren" wurde, ein wenig attraktives Bild bietet.

Das eingesetzte, sichtlich mit grossem Experimentieraufwand entwickelte (jedoch aufgrund der publizierten Literatur kaum vollständig zu durchschauende) Klassifizierungsverfahren ist in linguistischer Hinsicht aufwendiger als die im Rahmen von *DESIRE* und *WWLib* verwendeten Ansätze. Allerdings handelt es sich auch hier um kein System, das mit Trainingsdokumenten bzw. einem lernenden Klassifikator arbeitete. Leider fehlt auch eine systematische Evaluierung der Leistung des erstellten Klassifikators. Das Problem der Homonymie führte offensichtlich auch in diesem Projekt zu vielen Fehlklassifizierungen.

Ohne Zweifel ist *GERHARD* als bedeutendes Unternehmen einzuschätzen, im Rahmen dessen – im Gegensatz zu den Projekten in Lund – mit einer Universalklassifikation *und* – im Gegensatz zu dem Projekt in Wolverhampton – mit wirklich grossen Datenmengen gearbeitet wurde. Für den deutschen Sprachraum war und ist es das bisher einzige grosse Projekt zum automatischen Klassifizieren. Die Beschränkung auf deutsche Web-Ressourcen kann aber auch als ein grundsätzliches Problem des Systems *GERHARD* gesehen werden, da es m.E. doch etwas fraglich ist, ob wissenschaftliche Informationen nur aus diesem geographischen Bereich heutige Benutzerbedürfnisse abdecken (vgl. auch Krüger 1999, 46). Für das in diesem Zusammenhang überdies aufgetretene Harvesting-Problem (Einbeziehung zahlreicher nichtwissenschaftlicher Webseiten) hat etwa Tröger (1998, 1928–1929) die intellektuelle Kontrolle des Inputs nach dem Muster des Projekts *IBIS* (einer kooperativ gefüllten Datenbank) vorgeschlagen; die Chancen auf die (dauerhafte) Realisierung eines solchen Ansatzes sind m.E. jedoch wohl minimal.

Das Nachfolgeprojekt GERHARD II. Das nach wie vor in Entwicklung befindliche Folgeprojekt *GERHARD II* kann natürlich nur sehr vorsichtig beurteilt werden. Festzustehen scheint, dass dabei die grundsätzlichen Harvesting-Probleme durch den Wechsel der Robot-Software gelöst werden konnten. Ob bzw. wie es gelungen ist, die wissenschaftlich nicht relevanten Web-Ressourcen auszufiltern, ist zur Zeit noch nicht absehbar. Ziemlich sicher ist des weiteren, dass eine modernere und leistungsfähigere Benutzeroberfläche zur Verfügung stehen wird.

GERHARD II ist deshalb besonders interessant, da hier zum ersten Mal bei einer *grossen* Web-Anwendung ein auf dem maschinellen Lernen basierendes Verfahren eingesetzt wird. Die Erfahrungen aus diesem für die Projektmitarbeiter vermutlich nicht einfachen Vorgehen beim automatischen Klassifizieren könnten von grossem Gewinn für weitere Projekte dieser Art, zumal im deutschsprachigen Raum, sein. Sollte es gelingen, mit diesem Verfahren – auf der Basis einer methodisch einwandfreien und dokumentierten Evaluierung – eine zufriedenstellende Klassifizierungsgüte zu erzielen, so wäre dies wohl auch einer der ersten abgesicherten Hinweise darauf, dass sich der gegenwärtig als "state-of-the-art" betrachtete methodische Ansatz in einer praktischen Anwendung *mit sehr grossen Datenmengen* zu bewähren vermag.

8.5 Scorpion / OCLC

Koch et al. (1997, 34) bezeichneten *Scorpion* als "the most important project in the area of automatic classification." Diese Bedeutung ist m.E. vor allem darin zu sehen, dass eine grosse Institution wie OCLC sich des Themenkomplexes "automatisches Klassifizieren von Web-Dokumenten" annahm und dafür vermutlich nicht unerhebliche Ressourcen aufwandte.

Die im Rahmen von *Scorpion* eingesetzte Klassifizierungstechnik (ein traditioneller IR-Ansatz) selbst war wenig spektakulär und konnte – soweit dies aus der Projektliteratur hervorgeht – auch nicht durch besonders eindrucksvolle Ergebnisse überzeugen. Beeindruckend sind dagegen die vielfältigen Versuche zur Erweiterung des Vokabulars der DDC, die auch ahnen lassen, welche Möglichkeiten zur Vernetzung und Auswertung bestehender Informationsressourcen (Klassifikationen, Normdateien, bibliographische Datenbanken) einem führenden Infrastrukturunternehmen wie OCLC zur Verfügung stehen. Auch die Entwicklung bzw. der Einsatz von Werkzeugen wie jenen aus dem *WordSmith*-Projekt – sowohl für die Anreicherung der Dewey-Datenbank als auch die Aufbereitung der zu klassifizierenden Dokumente – ist in diesem Zusammenhang zu erwähnen. Von Interesse sind schliesslich auch manche der in Bezug auf die Ergebnisbewertung angestellten modellhaften Überlegungen.

Bedauerlich ist dagegen, dass nie eine mit *Scorpion* klassifizierte Dokumentenkollektion für ein DDC-basiertes Browsing öffentlich verfügbar gemacht wurde und dass die Projektliteratur den Eindruck vermittelt, dass viele Untersuchungsansätze unfertig verblieben sind oder zumindest allfällige Ergebnisse nicht dokumentiert wurden. Dies betrifft sowohl die bei OCLC angestellten Tests als auch die in Kooperation mit NetLab begonnenen Untersuchungen. Da es unwahrscheinlich ist, dass gute Resultate unveröffentlicht geblieben wären, lässt dies wohl auch Rückschlüsse auf die erzielten Klassifizierungsergebnisse zu.

8.6 Weitere Anwendungen und Projekte

Bücher. Ein etwas überraschendes Ergebnis der vorliegenden Arbeit ist das bislang nahezu völlige Fehlen jeglichen Interesses an einer automatischen Klassifizierung von Buchbeständen. Mag dies für den amerikanischen Raum noch dadurch erklärbar sein, dass neue MARC-Datensätze in der Regel bereits klassifiziert und auch bei retrospektiven Erfassungsprojekten (aufgrund der grossen Freihandaufstellungen) häufig Aufstellungsnotationen verfügbar sind, so besteht z.B. im deutschen Sprachraum eine andere Situation. In den 1990er Jahren wurde errechnet, dass allein in Deutschland 52 Millionen älterer (Formal-)Katalogisate auf ihre Konvertierung warteten (Beyersdorff 1993), von denen inzwischen ein gewisser Teil maschinell erfasst, aber vermutlich *nicht* sachlich (klassifikatorisch) erschlossen sein dürfte.

Aufgrund der für diese Arbeit durchgeführten Recherchen scheint festzustehen, dass mit Ausnahme der inzwischen schon mehr als ein Jahrzehnt zurückliegenden Untersuchung von Larson (1992) keine *signifikanten* Studien oder Anwendungen aus dem Bibliotheksbereich vorliegen, die sich mit dem automatischen Klassifizieren von Büchern bzw. Katalogisaten beschäftigen. Enttäuschend ist in diesem Zusammenhang v.a. auch, dass sich nicht einmal das bibliothekarische Infrastrukturunternehmen OCLC dieses Themas angenommen hat. Das von OCLC durchgeführte Projekt *Scorpion* zielte ausschliesslich auf Web-Dokumente ab; auch der im Rahmen von *CORC / Connexion* verfügbare Klassifikator dient nur der Katalogisierung elektronischer Ressourcen (vgl. Abschnitt 6.9.1).

So kann die von Gödert (2002a, 396) geäusserte Hoffnung, wonach die wachsende Zahl der in OPACs verzeichneten, aber nicht inhaltlich erschlossenen Bücher ein verstärktes Interesse an der automatischen Zuteilung von Notationen hervorrufen würde, zumindest auf der Basis der hier untersuchten Literatur nicht bestätigt werden.

Im Vergleich zur automatischen Klassifizierung von elektronischen Dokumenten, die ja zumeist im Volltext vorliegen, mutet die im Fall von Katalogisaten auf die Wörter aus Titel, Untertitel und (allfällig) verbaler Sacherschliessung begrenzte Textmenge sehr spärlich an. Dieser Umstand mag für Interessenten an automatischen Klassifizierungsverfahren abschreckend gewirkt haben. In diesem Zusammenhang sei auch an den Befund von Larson (1992) erinnert, wonach das sonst so bewährte TFIDF-Gewichtungsverfahren bei Katalogisaten besonders schlechte Ergebnisse erbracht habe. Zwar hat sich z.B. im Rahmen der *MILOS*-Projekte gezeigt, dass sogar auf dieser limitierten textuellen Basis mit einer linguistischen Methode des automatischen Indexierens eine sinnvolle Erweiterung des inhaltstragenden Wortschatzes erreicht werden kann (Sachse et al. 1998), doch steht der Beweis noch aus, dass dies auch für das automatische Klassifizieren gilt.

Patentliteratur / Mediendokumentation. Die mit Patentliteratur, v.a. Patentanmeldungen, sowie mit Pressetexten vorgenommenen Experimente und Tests haben gezeigt, dass in diesen beiden Bereichen grosses Interesse an einer Automatisierung des ressourcenintensiven Klassifizierungsprozesses besteht. Immerhin finden sich dort bereits zumindest vereinzelt Systeme, die im operativen Alltagsbetrieb angewandt werden. In beiden Bereichen sind aber auch die Grenzen erkennbar, die automatischen Klassifizierungsverfahren heute noch anhaften, zumal diese bisher nur als *semi-automatische* Lösungen eingesetzt werden. Für die sowohl in Patentämtern als auch in Pressearchiven erforderliche hohe Genauigkeit der Klassifizierung können auch die gegenwärtig besten Verfahren offensichtlich noch nicht garantieren.

Web-Portale, Suchmaschinen, Informationsdienste. Auch die bei *Lexis-Nexis, Northern Light* und *Factiva* eingesetzten Verfahren sind interessant, da es sich dabei um Anwendungen für den Produktionsbetrieb handelt. Auch wenn die betreffenden Firmen hinsichtlich der konkret angewandten Methodik wohl nicht alle Details aufdecken, geht aus den verfügbaren Beschreibungen hervor, dass dabei *regelbasierte* Verfahren immer noch eine Rolle spielen, aber auch schon moderne, auf dem maschinellen Lernen beruhende Klassifikatoren eingesetzt werden. Die jeweils erzielte Klassifizierungsgüte ist der vorliegenden Literatur zufolge auffallend hoch.

Nicht so erfolgreich fiel die maschinelle Klassifizierung der virtuellen Bibliothek *INFOMINE* aus, bei der zwar die LCC und ein modernes Verfahren eingesetzt wurden, die erzielte Klassifizierungsgüte jedoch schon nach der obersten Hierarchiestufe deutlich absank. Angesichts der geringen Grösse dieser Datenbank mag eine grobe Kategorisierung vielleicht ausreichend gewesen sein; überträgt man das Ergebnis jedoch auf eine wirklich grosse Datensammlung, so könnte damit günstigstenfalls eine Rangreihung von Suchergebnissen innerhalb einer Hauptklasse, nicht jedoch eine zufriedenstellende Browsing-Struktur erreicht werden. Als interessant beim Ansatz von *INFOMINE* ist zu der Umstand werten, dass dabei für das Training Datensätze aus einem Bibliothekskatalog herangezogen wurden – eine Parallele zu den Projekten *GERHARD II* und *Pharos* (vgl. Abschnitte 5.3 und 7.4.5). Diese Funktion von Bibliothekskatalogen, als Trainingskollektionen für Klassifikatoren zu dienen, die sodann für die Erschliessung elektronischer Dokumentenkollektionen eingesetzt werden, ist ein durchaus bemerkenswertes Novum.

Die weiteren referierten Anwendungen und Projekte weisen auf die grosse Vielfalt einschlägiger Vorhaben, von denen es sicherlich noch eine Reihe weiterer gibt, hin. In methodischer Hinsicht ist dabei eindeutig eine Ausrichtung auf lernende Klassifikatoren zu erkennen. Mit Ausnahme von *eBay*, wo enorme Datenmengen umgesetzt und erschlossen werden, handelt es sich allerdings nicht um sehr grosse Projekte. Zu *eBay* wäre noch zu bemerken, dass die dort anfallenden Datensätze im Hinblick auf die Kürze

der inhaltstragenden Texte eine gewisse Parallelität zu bibliothekarischen Datensätzen aufweisen, möglicherweise aber durch ein weniger variationsreiches Vokabular gekennzeichnet sind, da sich die Beschreibungen der Auktionswaren vermutlich mehr gleichen als die Titel von Büchern in Online-Katalogen.

8.7 Andere Aspekte

An dieser Stelle sollen noch kurz einige Aspekte diskutiert werden, die in der Literatur zum automatischen Klassifizieren immer wieder angesprochen werden und im Zusammenhang mit konkreten Anwendungen und Projekten nicht ausreichend behandelt werden konnten.

Sprachen / Mehrsprachigkeit. Der bei weitem grösste Teil der Literatur zum automatischen Klassifizieren stammt aus dem anglo-amerikanischen Sprachraum und bezieht sich – oft sogar implizit, d.h. ohne es überhaupt zu reflektieren – auf Anwendungen und Projekte in englischer Sprache. Daher erschöpft sich in diesen Arbeiten der linguistische Methodenaspekt meist in Varianten des Stemming; oft wurde auf eine linguistische Bearbeitung der Attribute sogar gänzlich verzichtet.

Neben der automatischen Klassifizierung englischsprachiger Texte wurde auch über Anwendungen auf Dokumente in deutscher, flämischer, französischer, japanischer, portugiesischer und spanischer Sprache berichtet. Was deutschsprachige Texte betrifft, so wurde im Rahmen des Projektes *GERHARD* mehr Wert auf die linguistische Vorbereitung der Texte gelegt als sonst meist üblich, während etwa in der Pressedokumentation von Gruner + Jahr wiederum nur Stemming zur Anwendung kommt.

Mehrsprachige Anwendungen sind bislang die Ausnahme geblieben. Die bedeutendste davon war *GERHARD* (Klassifizierung in deutscher und englischer Sprache, Suchoberfläche zusätzlich in französischer Sprache). Die Experimente im Rahmen des Projekts *PEKING* (Englisch, Spanisch) wurden oben kurz erwähnt, ebenso die Absichten des Projekts GRACE (Englisch, Deutsch, Schwedisch, Italienisch). Die potentiell mehrsprachigen Anforderungen in der Patentdokumentation (z.B. Europäisches Patentamt, WIPO, INPI) wurden bislang nicht realisiert.

In einem Experiment zeigten kürzlich Lauser & Hotho (2003), dass Support Vector Maschinen nahezu perfekt zwischen den drei Sprachen Englisch, Französisch und Spanisch zu unterscheiden vermochten, wenn diese als "Klassen" mit einem entsprechenden dreisprachigen Textkorpus trainiert wurden. Dieses Verfahren bietet sich somit für den Einsatz bei der automatischen Spracherkennung an. Des weiteren versuchten diese Autoren eine mehrsprachige Klassifizierung mittels SVM, wobei die "Klassen" Deskriptoren aus dem AGROVOC-Thesaurus waren. Dabei manipulierte man die mit englischsprachigen Trainingsdokumenten erstellten Klassenvektoren so,

dass nur die in einer auf der Basis des genannten Thesaurus erstellten fünfsprachigen fachspezifischen Ontologie aufscheinenden Terme Verwendung fanden. Damit konnten auch Dokumente in den vier anderen Sprachen mit nahezu gleicher Güte klassifiziert werden wie die englischsprachigen.

Die von einer Reihe kommerzieller Anbieter von Klassifizierungssoftware (z.B. Convera 2002; Inxight 2004; Kofax 2002; Recommind o.J.) behauptete Eignung ihrer Produkte für mehrsprachige Anwendungen – aufgrund mehrsprachiger Versionen, Wörterbücher, sprachunabhängiger Mustererkennungsverfahren – müsste im einzelnen kritisch hinterfragt bzw. geprüft werden.

Klassifikationsbasierte Benutzeroberflächen. Ein mehrsprachiger Zugriff zu einer klassifikatorisch erschlossenen Kollektion kann auch ausschliesslich Sache der Benutzerschnittstelle sein, wie bspw. "Französisch" im Fall von *GERHARD*, wo nur eine französischsprachige Version des Klassifikationssystems vorlag, aber keine Klassifizierung französischer Dokumente damit erfolgte. Der Themenaspekt "Benutzerinterface" stellt naheliegenderweise eine mit dem automatischen Klassifizieren elektronischer Dokumente nahe verwandte Thematik dar. Obzwar nicht im Detail Gegenstand dieser Arbeit, wurde bei der Darstellung der vier grossen und z.T. auch der anderen hier analysierten Projekte zumindest überblicksartig auch auf die Benutzungsmöglichkeiten und die jeweils gewählte Benutzeroberfläche eingegangen.

Die Gestaltung von klassifikationsbasierten Benutzeroberflächen in Datenbanken elektronischer Dokumente (insbesondere Web-Ressourcen) oder auch in bibliothekarischen Online-Katalogen ist ein komplexes Thema, das hier nicht näher behandelt werden kann und das sicherlich eine eigenständige Übersichtsarbeit rechtfertigen würde. Von den traditionellen systematischen Bibliothekskatalogen (Zettelkatalogen) ist bekannt, dass diese – sofern eine Alternative in Form eines Schlagwortkatalogs vorhanden war – wenig bzw. inadäquat genutzt wurden. Ähnliches gilt für die vielfach völlig uninspirierte Gestaltung der klassifikationsbasierten Recherche in OPACs (Wheatley 2000, 121f.). Dass dies auch anders möglich ist, haben die (relativen) Erfolge von hierarchisch organisierten Web-Katalogen wie *Yahoo!*, *LookSmart* oder *Google* oder auch von OPAC-Systemen wie *Book House* (Pejtersen 1993) gezeigt. Spätestens seit der Arbeit von Allen (1995), der Interfaces für ein traditionelles, hierarchisches Klassifikationssystem (DDC) sowie ein facettiertes System (*ACM Computing Reviews*) gestaltete, sind derartige Oberflächen ein Thema, das bis heute aktuell ist (vgl. z.B. Gödert 2002b). Wohl zu Recht wurde freilich der Mangel an empirischen Überprüfungen der Benutzbarkeit solcher Interfaces – etwa im Gegensatz zu der für viele Benutzer erwiesenermassen problematischen Booleschen Recherche – kritisiert (Hobohm 1998, 9).

Klassifizierung von Ergebnismengen vs. Dokumenten. Verwandt mit dem Aspekt "Benutzeroberfläche" ist der Ansatz, nicht die gesamte Dokumentenkollektion, sondern bloss die jeweils aktuellen (stichwortbasierten) Suchresultate zu klassifizieren und diese dem Benutzer entweder in einer hierarchischen Anzeige oder zumindest in einer mit der Bezeichnung der jeweiligen Klasse(n) angereicherten Kurzanzeige zu präsentieren. Für das letztere Modell fanden z.B. Drori & Alon (2003) bei einer empirischen Untersuchung heraus, dass diese Form der Anzeige bei den befragten Benutzern besser ankam als die übliche von Suchmaschinen gezeigte Ergebnisliste. Oft basieren diese Ansätze jedoch nicht auf dem automatischen Klassifizieren, sondern auf *Clustering*, d.h., die Klassen werden induktiv auf der Basis der Ergebnisdokumente erstellt und nach einem heuristischen Prinzip benannt. So führt z.B. das Interface *Grouper* die Benennung durch die Ermittlung von "shared phrases", d.h. der am häufigsten auftretenden Wörter und Phrasen in den Dokumenten der einzelnen Cluster durch (Zamir & Etzioni 1998). Eine Reihe von Suchmaschinen bietet ähnliche Gruppierungen von Ergebnismengen an (Calishain 2002a; 2002b; Topic Clustering 2002).

Ein klassifikationsbasierter Ansatz ging dagegen vom Klassifikationsschema des Web-Katalogs *LookSmart* aus (Chen & Dumais 2000; Dumais & Chen 2000). Dabei wurde ein statistisches Klassifikationsmodell (SVM) offline trainiert und bei jeder Suche "on-the-fly" auf die (stichwortbasierte) Ergebnismenge angewandt. Im Gegensatz zum oben erwähnten clusteranalytischen Ansatz hat diese Vorgangsweise den Vorteil, den Benutzern konsistente und eventuell sogar bekannte Klassenstrukturen darzubieten. "Run time classification" erwies sich als sehr effizient (im Vergleich zum Clustern). Bei der Ergebnisklassifizierung zeigte sich auch, dass eine Beschränkung auf die obersten Ebenen der Hierarchie ausreichte, um den Benutzern eine adäquate Disambiguierung der Ergebnismenge zu bieten.

Metadaten. Während etwa im Rahmen des Projektes *DESIRE II* den im Dokument enthaltenen Metadaten das höchste Gewicht zugeteilt wurde (vgl. Abschnitt 3.2.3), vertritt Sebastiani (2002a; 2002b; 2004) den Standpunkt, dass ein modernes, auf maschinellem Lernen basierendes Verfahren die Klassifizierungsaufgabe gänzlich *ohne* exogene Informationen – dazu zählen auch Metadaten – bewältigen solle (s. a. Abschnitt 2.1.1).

Dies sind nur scheinbar kontroverse Ansichten, denn auch bei modernen Verfahren hat sich bestätigt, dass diese auf der Basis von Metadaten bessere Precision- und Recall-Werte erzielen als auf der Basis des Textes der Dokumente (z.B. Pierre 2001, 8). Die Aussage Sebastianis ist wohl nicht als Aufforderung zum Verzicht auf die Verwendung von Metadaten beim automatischen Klassifizieren zu verstehen, sondern eine angesichts der Realsituation am WWW wichtige Forderung an die Methodik. Realistischerweise muss davon ausgegangen werden, dass (derzeit) standardisierte Metadatenstrukturen nur geringfügig verbreitet sind. Zudem ist "Metadaten" nicht zwangsläufig

mit "inhaltsbeschreibende Metadaten" gleichzusetzen. So stellte sich etwa im Rahmen des Projekts *GERHARD* heraus, dass lediglich 0,11% der gesammelten HTML-Seiten Metadaten nach Dublin Core enthielten (Wätjen et al. 1998, 5). Zu einem ähnlichen Resultat gelangte man bei *Northern Light*: Zwar wurde anhand von zwei extern durchgeführten Studien festgestellt, dass 30–40% aller Webseiten Metadaten beinhalteten, doch handelte es sich dabei grösstenteils um von HTML-Editoren ("page creation software") produzierte Meta-Tags ohne jegliche thematische Information. Sofern sachliche Metadaten vorhanden waren, handelte es sich zumeist um "spam" (d.h. eine vom Betreiber der betreffenden Webseite bewusst vorgenommene, unverhältnismässige Anhäufung von Termen zum Zweck der Indexierung durch Suchmaschinen). Der Anteil wohlgeformter Metadaten wurde dagegen durch die Metapher "Spurenelemente" charakterisiert (Ward, 1999). Bei einer Analyse von ca. 20.000 amerikanischen Webseiten fand Pierre zwar heraus, dass immerhin knapp ein Drittel dieser Seiten Information in den Meta-Feldern "keywords" oder "description" (von mehr als null Wörtern) aufwies, meinte aber im Hinblick auf das gesamte Web dennoch, dass die Verwendung von Metadaten "inconsistent or non-existent" sei (2000, 2).

Im Hinblick auf das automatische Klassifizieren stellt sich daher eher die Frage, *wie damit* zur Anreicherung von Web-Dokumenten mit inhaltskennzeichnenden Metadaten beigetragen werden könne. Überlegungen dieser Art gab es u.a. im Rahmen der Projekte *WWLib* (Jenkins et al. 1999) und *Scorpion* (Shafer 1997b). Ein Ansatz, der einen Klassifikator mit einer Technik zur Übertragung von vorhandener Meta-Information auf verlinkte Seiten kombiniert ("back-propagation"), wurde von Marchiori (1998) vorgeschlagen.

Welches Klassifikationssystem? Im Zusammenhang mit dem Klassifizieren von Web-Ressourcen wird immer wieder diskutiert, welches Klassifikationssystem bzw. welche Art von Klassifikationssystemen sich dafür am besten eignen würde. Da diese Frage auch in der Literatur zum automatischen Klassifizieren behandelt wird (obwohl sie eigentlich davon unabhängig zu betrachten ist), soll abschliessend auch hier kurz darauf eingegangen werden.

Argumente für die Anwendung *traditioneller bibliothekarischer Klassifikationssysteme* wie DDC, UDC und LCC werden vor allem von bibliothekarischer Seite (Koch 1998c; Koch et al. 1999; Mitchell & Vizine-Goetz 2002; Vizine-Goetz 1999b; 2002) geäussert; es handelt sich um wohlbekannte Kriterien, die hier zu wiederholen nicht erforderlich ist. Für die Verwendung der neu entstandenen *Web-Klassifikationen* (von Katalogdiensten wie *Yahoo!, DMOZ, LookSmart* u.a.) wird z.B. ins Treffen geführt, dass die fachliche Verteilung von Web-Ressourcen völlig anders sei als jene von gedrucktem Material (Jenkins & Inman 2000,506), dass diese Schemata die moderne und sich verändernde Welt widerspiegeln (Koch et al. 1999, Kpt. 7) oder dass die Benutzer die Fra-

ge durch ihre eindeutige Präferenz dieser neuen Klassifikationssysteme längst entschieden hätten (Wheatley 2000, 124).

Von bibliothekarischer Seite werden diesen neu entstandenen Systemen jedoch grobe Schwächen attestiert: Fehler in Logik und Hierarchie, Inkonsistenz bei der Definition und Anordnung der Klassen, mangelnde Spezifität, fehlende bzw. fehlerhafte relationale Struktur, Vermischung unterschiedlicher Klassifikationskriterien und Dimensionen (Godby & Vizine-Goetz 2000, 22; Koch et al. 1999c; Kwasnik & Liu 2000; Zins 2002). Es hat aber auch den Anschein, als bemühe man sich angestrengt um den Beweis, dass traditionelle Schemata (z.B. die DDC) den neu entwickelten (z.B. jenem von *Yahoo!*) im Prinzip gar nicht so unähnlich seien (Vizine-Goetz 2002).

Kritik anderer Art kommt hingegen von den Befürwortern *facettierter Klassifikationssysteme*. So wirft etwa Bates (2002, 3) den Anwendern hierarchischer Schemata – traditioneller wie neu entwickelter – vor, "altmodische" Systeme anzuwenden ("faceted classification is to hierarchical classification as relational databases are to hierarchical databases"). Ähnlich argumentierte bereits Woodward (1996, 192) mit Bezug auf das Schema von *Yahoo!* ("It seems that in reinventing the wheel, Yahoo! has stepped back in time, not moved forward"). Grössere Flexibilität, bessere kombinatorische Möglichkeiten (sowohl im Hinblick auf Spezifität als auch Exhaustivität), grössere Hospitalität (gegenüber neuen Themen), bessere Möglichkeiten zur Repräsentation komplexer Beziehungen zwischen Fachgebieten sind die den facettierten Systemen (z.B. Bliss, Colon) zugeschriebenen Vorteile (Broughton & Lane 2000; Ellis & Vasconcelos 1999; Patel 2002). Der Nachteil des Fehlens einer Browsing-Hierarchie könne z.B. durch ein interaktives Benutzerinterface wieder ausgeglichen werden (Patel 2002, 8). Während die Verwendung von Facettenklassifikationen für die Erschliessung von (wissenschaftlichen) Web-Ressourcen bisher jedoch kaum realisiert ist, weist eine rezente Untersuchung darauf hin, dass im Bereich des "e-commerce" der Einsatz (einfacher) facettierter Systeme bereits überwiegen dürfte (Adkisson 2003).

8.8 Ausblick

"Soon, every electronic resource will be run through Scorpion or other automatic classification tools" (Shafer 1997c, sl. 20). Diese auf dem Höhepunkt der Entwicklung der vier grossen Projekte zum automatischen Klassifizieren geäusserte Prognose hat sich als viel zu optimistisch erwiesen. In der Tat hat es den Anschein, als sei der Elan, mit dem diese Projekte damals begonnen und teilweise umgesetzt wurden, im neuen Jahrtausend wieder zum Erliegen gekommen. Über die dafür massgeblichen Gründe – zu geringer Klassifizierungserfolg, Interessensverlagerung bei den Beteiligten, mangelnde finanzielle Ressourcen usw. – soll hier nicht weiter spekuliert werden. Faktum ist jedenfalls,

dass drei dieser Projekte zur Zeit "eingefroren" sind und das vierte (*Scorpion*) im wesentlichen nur hinter verschlossenen Türen, d.h. für registrierte Kunden nutzbar ist. Das in Entwicklung befindliche Projekt *GERHARD II* bietet wenigstens einen gewissen Hoffnungsschimmer im Hinblick auf eine grossflächige automatische Erschliessung elektronischer Ressourcen.

Das geringe Interesse an der automatischen Klassifizierung von *Büchern* wurde bereits weiter oben konstatiert bzw. beklagt. Diesbezüglich besteht jedoch aufgrund des Übersetzungsprojekts *DDC Deutsch* (Gödert 2002a) zumindest die Chance, dass im Falle der Akzeptanz der DDC als "Einheitsklassifikation" für deutschsprachige Online-Kataloge auch Bestrebungen zur Ausstattung eines möglichst grossen Teils der darin enthaltenen Katalogisate mit Notationen aus diesem System entstehen könnten. Hier soll nicht weiter über die verschiedenen Möglichkeiten zur Übernahme solcher Notationen diskutiert werden; für eine automatische Klassifizierung stünden aber sicherlich Trainingsdokumente in grosser Zahl zur Verfügung, wenngleich der Anteil deutschsprachiger Katalogisate unter den weltweit nach der DDC erschlossenen Titel wohl nur gering sein dürfte. Ein anderes Klassifikationssystem, für das grosse Zahlen bereits erschlossener Katalogisate – darunter auch sehr viele deutschsprachige – existieren, ist die *Basisklassifikation,* die zudem in drei Sprachen (niederländisch, englisch, deutsch) vorliegt. Mit diesem System, das für bestimmte Funktionen in Bibliothekskatalogen durchaus sehr gut geeignet ist, zu experimentieren und dabei auch wichtige Erfahrungen in methodischer Hinsicht zu gewinnen, würde sich gewiss lohnen.

Treibende Kraft beim automatischen Klassifizieren grosser Datenmengen kann heute am ehesten den Bereichen Patentdokumentation und Mediendokumentation attestiert werden. Sowohl im Hinblick auf eine grossflächige automatische Erschliessung von Web-Ressourcen als auch aus bibliothekarischer Sicht sollten daher die Entwicklungen auf diesen Gebieten, die vielleicht nicht kurzfristig, aber sicherlich mittelfristig auf eine qualitativ zufriedenstellende Vollautomatisierung abzielen, genau verfolgt werden. Ähnliches gilt auch für (kommerzielle) Web-Portale und Informationsanbieter, doch müssen die Chancen, von dieser Seite detaillierte Informationen über Verfahren bzw. methodische Vorgangsweisen zu erhalten, als geringer eingeschätzt werden. Eine weitere treibende Kraft auf dem kommerziellen Sektor stellen sicherlich die in jüngerer Vergangenheit immer mehr auf den Markt drängenden Software-Tools für das automatische Klassifizieren dar. Wie sich in der vorliegenden Studie herausgestellt hat, haben Produkte dieser Art – wenn auch bislang nur in wenigen Fällen – bereits Eingang in den wissenschaftlichen bzw. nichtkommerziellen Anwendungsbereich gefunden.

Schliesslich zeigt die grosse Zahl verschiedener sonstiger Projekte und Anwendungen, von denen viele aufgrund ihres primär methodologischen Charakters im Rahmen dieser Arbeit gar nicht behandelt wurden, dass grundsätzlich ein sehr weitgefächer-

tes Interesse an der automatischen Textklassifizierung besteht, das die oben geäusserte Skepsis ein wenig relativiert. Eine wichtige Rolle kommt dabei jedenfalls auch den zahlreichen Autoren und Experimentatoren aus unserer Nachbardisziplin *Informatik* zu, deren Beiträge allerdings für Rezipienten aus dem Bereich der Informationswissenschaft oft schwer überwindliche Barrieren aufweisen. Im Rahmen dieser Arbeit hat sich jedoch gezeigt, dass die Anstrengung, über die Grenzen der eigenen Disziplin hinauszublicken, sicherlich lohnend sein kann.

Literaturverzeichnis

Aagaard 1995
Aagaard, P. (1995). Tværgående emnesøgninger i fælleskatalogen DANBIB: Automatisk konstruktion af ækvivalenstabeller mellem forskellige klassifikationssystemers notationer [Text in dänischer Sprache].[1] *Tidskrift for dokumentation*. 50(1). 21–31.

Adams 2003
Adams, K. C. (2003). Word wranglers: Automatic classification tools transform enterprise documents from "bags of words" into knowledge resources. *IntelligentKM*. WWW, 5p. <http://www.intelligentkm.com/feature/010101/feat1.shtml> [03.05.2003]

Adkisson 2003
Adkisson, H. P. (2003). *Use of faceted classification*. WWW, 3p. <http://www.webdesignpractices.com/navigation/facets.html> [16.05.2004]

Allen 1995
Allen, R. B. (1995). Two digital library interfaces that exploit hierarchical structure. [Vortrag:] *DAGS95: Electronic Publishing and the Information Superhighway, Boston, MA, May 30 – June 2, 1995*. WWW, 13p. <http://raven.umd.edu/~rba/PAPERS/LIBR/libr.html> [22.02.2004]

AmikaNow! 2002a
AmikaNow! Corporation <Kanata, Ontario, CA> (2002). *AmikaClassifier*. WWW, 2p. <http://www.amikanow.com/products/classifier.asp> [10.03.2004]

AmikaNow! 2002b
AmikaNow! Corporation <Kanata, Ontario, CA> (2002). *AmikaClassifier SDK*. WWW, 2p. <http://www.amikanow.com/products/PDF/ANClassifier_glossy2.pdf> [29.04.2004]

Ardö & Koch 1993
Ardö, A.; Koch, T. (1993). Wide-Area Information Server (WAIS) as the hub of an electronic library service at Lund University. *Opportunity 2000: Understanding and serving users in an electronic library; 15th Internat. Essen Symposium, 12–15 Oct. 1992*. Hrsg.: Helal, A. H. et al. – Essen: Univ.-Bibliothek Essen. 199–210.

Ardö & Koch 1994
Ardö, A.; Koch, T. (1994). *Nordic WAIS/World Wide Web Project: Subproject: Automatic classification of WAIS databases.* WWW, 3p. <http://www.lub.lu.se/autoclass.html> [02.03.2004]

Ardö & Koch 1999a
Ardö, A.; Koch, T. (1999). Automatic classification applied to full-text Internet documents in a robot-generated subject index. *Online information 99: Proceedings; 23rd International Online Information Meeting, London, 7–9 Dec. 1999*. Hrsg.: McKenna, B.; Graham, C. – Oxford: Learned Information International. 239–246.

Ardö & Koch 1999b
Ardö, A.; Koch, T. (1999). Creation and automatic classification of a robot-generated subject index. *ACM Digital Libraries '99: Proceedings of the 4th ACM Conference on Digital Libraries, Berkeley, CA, Aug. 11–14, 1999*. Hrsg.: Fox, E. A.; Rowe, N. – New York, NY: ACM Press 210–211. – Prepr.: WWW, 4p. <http://www.lub.lu.se/desire/poster.html> [22.02.2004]

[1] Titelübersetzung: Recall improved subject searches in the national union catalog DANBIB: Automatic establishing of equivalence tables between different classification schemes [Originaltext].

Ardö & Koch 2000
Ardö, A.; Koch, T. (2000). *Automatic classification demonstration page (DESIRE II)*. WWW, 3p. L/U: 2000-02-24. <http://www.lub.lu.se/desire/demonstration.html> [15.05.2004]

Ardö & Lundberg 1998
Ardö, A.; Lundberg, S. (1998). A regional distributed WWW search and indexing service: The DESIRE way. [Vortrag:] *7th International World-Wide Web Conference, April 14–18, 1998, Brisbane, Australia.* WWW, 13p. <http://decweb.ethz.ch/WWW7/1900/com1900.htm> [03.05.2004]

Ardö et al. 1994a
Ardö, A.; Falcoz, F.; Koch, T.; Nielsen; M.; Sandfær, M. (1994). *W4: Nordic WAIS/World Wide Web Project – project description and plan.* WWW, 13p. May 1993 (revised Feb. 1994). <http://www.lub.lu.se/W4/plan.html> [10.03.2004]

Ardö et al. 1994b
Ardö, A.; Falcoz, F.; Koch, T.; Nielsen; M.; Sandfær, M. (1994). Improving resource discovery and retrieval on the Internet: The Nordic WAIS/World Wide Web Project – summary report. *NORDINFO-Nytt.* (4). WWW, 11p.
<http://www.nordinfo.helsinki.fi/publications/nordnytt/nnytt4_94/sandfaer.htm> [07.01.2004]
Prepr.: WWW, 14p. <http://www.lub.lu.se/W4/summary.html> [10.02.2004].

Ardö et al. 1999
Ardö, A.; Koch, T.; Noodén, L. (1999). *The construction of a robot-generated subject index: DESIRE II D3.6a, working paper 1*. WWW, 8p. L/U: 1999-03-01.
<http://www.lub.lu.se/desire/DESIRE36a-WP1.html> [10.04.2003)

Ardö et al. 2002
Ardö, A.; Lundberg, S.; Zettergren, A.-S. (2002). Another piece of cake.....? [On the creation and history of NetLab]. *Ariadne.* (32). WWW, 7p. <http://www.ariadne.ac.uk/issue32/netlab-history/> [06.01.2004]

Attardi et al. 1999
Attardi, G.; Gulli, A.; Sebastiani, F. (1999). Automatic Web page categorization by link and context analysis. *Proceedings of THAI-99, 1st European symposium on telematics, hypermedia and artificial intelligence, Varese, Italy, 1999.* Hrsg.: Hutchison, C.; Lanzarone, G. – o.O. 105–119. [Prepr.:] WWW, 16p. <http://faure.iei.pi.cnr.it/~fabrizio/Publications/THAI99.pdf> [24.02.2004]

Autonomy o.J.
Autonomy, Inc. <San Francisco, CA, US; Cambridge, UK> (o.J.). Automatic classification. WWW, 6p. <http://www.autonomy.com/Content/Products/IDOL/f/Classification#01> [10.04.2003]

Bates 1998
Bates, M. J. (1998). Indexing and access for digital libraries and the Internet: Human, database, and domain factors. *Journal of the American Society for Information Science*. 49(13). 1185–1205.

Bates 2002
Bates, M. (2002). After the dot-bomb: Getting Web information retrieval right this time. *First Monday*. 7(7). WWW, 9p. <http://firstmonday.org/issues/issue7_7/bates/> [11.07.2002]

Bel et al. 2003
Bel, N.; Koster, C. H. A.; Villegas, M. (2003). Cross-lingual text categorization. *Research and advanced technology for digital libraries: 7th European conference (ECDL 2003), Trondheim, Norway, August 17–22, 2003; Proceedings.* Hrsg.: Koch, T. et al. – Berlin u.a.: Springer. (Lecture notes in computer science; 2769). 126–139.

Beyersdorff 1993
Beyersdorff, G. (1993). Gesamtergebnisse und Empfehlungen (Kapitel 5). *Retrokonversion: Konversion von Zettelkatalogen in deutschen Hochschulbibliotheken: Methoden, Verfahren, Kosten.* Hrsg.: Weber, K. – Berlin: Deutsches Bibliotheksinstitut. 285–311.

Blumberg & Atre 2003
Blumberg, R.; Atre, S. (2003). Automatic classification: Moving to the mainstream. *DM review*. (April). WWW, 10p. <http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=6501> [20.11.2003]

Broughton & Lane 2000
Broughton, V.; Lane, H. (2000). Classification systems revisited: Applications to Web indexing and searching. *Journal of Internet cataloging*. 2(3/4). 143–155.

Brückner 2001
Brückner, T. (2001). Textklassifikation. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Hrsg.: Carstensen, K.-U. et al. – Heidelberg, Berlin: Spektrum. 442–447.

Burden 1997
Burden, P. (1997). *Design and construction of a search engine*. WWW, 7p. <http://www.scit.wlv.ac.uk/seed/docs/wwlibtng.html> [15.03.2004]

Burden 1998a
Burden, P. (1998). *The Automatic Classification Engine*. WWW, 9p. <http://seed.scit.wlv.ac.uk/docs/old.ace.html> [10.03.2004]

Burden 1998b
Burden, P. (1998). *Automatic classification results*. WWW, 7p. <http://seed.scit.wlv.ac.uk/docs/old.ace.results.html> [10.03.2004]

Burden 2000
Burden, J. P. H. (2000). *Stemming algorithms and their use*. WWW, 6p. <http://www.scit.wlv.ac.uk/seed/docs/mypapers/stemalg.html> [29.05.2004]

Burden & Jackson 1999
Burden, J. P. H.; Jackson, M. S. (1999). WWLib-TNG: New directions in search engine technology. [Vortrag:] *IEE Informatics Colloquium: Lost in the Web – Navigation on the Internet, Nov. 1999*. WWW, 9p.+18sl. <http://seed.scit.wlv.ac.uk/papers/iee.html> [10.03.2004] und <http://www.scit.wlv.ac.uk/~jphb/myppt/iee.ppt> [20.11.2003]

Calishain 2002a
Calishain, T. (2002). Clustering with search engines. Law library resource Xchange. (June 3). WWW, 3p. <http://www.llrx.com/features/clusteringsearch.htm> [12.12.2002]

Calishain 2002b
Calishain, T. (2002). Clustering with search engines, part 2. *Law library resource Xchange*. (June 17). WWW, 3p. <http://www.llrx.com/features/clusteringsearch2.htm> [12.12.2002]

Ceci & Malerba 2003
Ceci, M.; Malerba, D. (2003). Hierarchical classification of HTML documents with WebClassII. *Advances in information retrieval: 25th European conference on IR research (ECIR 2003), Pisa, Italy, April 14–16, 2003; Proceedings*. Hrsg.: Sebastiani, F. – Berlin u.a.: Springer. (Lecture notes in computer science; 2633). 57–72.

Chakrabarti et al. 1998
Chakrabarti, S.; Dom, B.; Agrawal, R.; Raghavan, P. (1998). Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *VLDB journal*. (7). 163–178.

Chekuri et al. 1997
Chekuri, C.; Goldwasser, M. H.; Raghavan, P.; Upfal, E. (1997). Web search using automatic classification. [Vortrag:] *6th international World Wide Web conference, Santa Clara, CA, April 1997*. WWW, 10p.
<http://theory.stanford.edu/people/wass/publications/Web_Search/Web_Search.html> [03.05.2003]

Chen & Dumais 2000
Chen, H.; Dumais, S. (2000). Bringing order to the Web: Automatically Categorizing Search Results. The future is here: Proceedings of CHI-00; ACM international conference on human factors in computing systems, The Hague, NL, April 1–5, 2000. Hrsg.: Turner, T. et al. – New York, NY: ACM Press. 145–152.

Cheng & Wu 1995
Cheng, P. T. K.; Wu, A. K. W. (1995). ACS: An automatic classification system. *Journal of information science.* 21(4). 289–299.

Chung & Noh 2003
Chung, Y. M.; Noh, Y.-H. (2003). Developing a specialized directory system by automatically classifying Web documents. *Journal of information science*. 29(2). 117–126.

Convera 2002
Convera AG Schweiz <Wil, CH> (2002). *Von der Information zum Wissen: RetrievalWare.* WWW, 7p. <http://www.convera.ch/Brochures/RetrievalWare.pdf> [07.10.2003]

Cremer & Neuroth 2002
Cremer, M.; Neuroth, H. (2002). Internationale Kooperation mit dem CORC-Projekt von OCLC an der SUB Göttingen. *Bibliothek: Forschung und Praxis.* 26(3). 261–271.

Cross 2000
Cross, P. (2000). Harvesting quality resources from the Web: The Social Science Search Engine. [Vortrag:] *DESIRE II Web Indexing Workshop, Delft, NL, 13–14 May, 2000*. WWW, 6p. <http://www.terena.nl/tech/projects/desire/d2-workshop/d2webindex2000/docs/phil.cross-index_wkshp_paper.doc> [18.02.2004]

Cross et al. 2000a
Cross, P.; Brickley, D.; Koch, T. (2000). *Enhancements to user interface: Deliverable D3.6b, DESIRE II*. WWW, 26p.<http://www.desire.org/html/research/deliverables/D3.6/> [20.11.2003]

Cross et al. 2000b
Cross, P.; Day, M.; Koch, T.; Peereboom, M.; Zettergren, A.-S. (2000). Subject classification, browsing and searching: [Chapter 2.5 of:] *DESIRE information gateway handbook.* WWW, 18p. <http://www.desire.org/handbook/2-5.html> [01.12.2003]

Davis et al. 2003
Davis, C.; Miller, K.; O'Shea, A.; Elder, C. (2003). *Classification of the Web*.WWW, ca. 35p. <http://www.slais.ubc.ca/courses/libr517/02-03-wt2/projects/classification/Index.htm> [01.12.2003]

Delphi Group 2002
Delphi Group <Boston, MA, US> (2002). *Taxonomy & content classification: Market milestone report; A Delphi Group white paper.* Boston, MA, Delphi Group. 60p. WWW:
<http://www.delphigroup.com/research/whitepapers/WP_2002_TAXONOMY.PDF> [03.05.2003]

Diekmann 2004
Diekmann, B. (2004). Telefongespräche mit dem Verfasser, 08./09.06.2004.

DIN 31623, Teil 1
Normenausschuss Bibliotheks- und Dokumentationswesen (NABD) im DIN Deutsches Institut für Normung e.V. <Berlin, DE> (1988). *Indexierung zur inhaltlichen Erschliessung von Dokumenten: Begriffe, Grundlagen Begriffe, Grundlagen (DIN 31 623, Teil 1).* Berlin, DIN.

DIN 32705
Normenausschuss Klassifikation (NAK) im DIN Deutsches Institut für Normung e.V. <Berlin, DE> (1987). *Klassifikationssysteme: Erstellung und Weiterentwicklung von Klassifikationssystemen (DIN 32 705)*. Berlin, DIN.

Dörre et al. 2001
Dörre, J.; Gerstl, P.; Seiffert, R. (2001). Volltextsuche und Text Mining. *Computerlinguistik und Sprachtechnologie: Eine Einführung.* Hrsg.: Carstensen, K.-U. et al. – Heidelberg, Berlin: Spektrum. 425–441.

Dolin et al. 1998
Dolin, R.; Agrawal, D.; El Abbadi, A.; Pearlman, J. (1998). Using automated classification for summarizing and selecting heterogeneous information sources. *D-Lib magazine.* (January). WWW, 11p.+10p.refs./apps. <http://www.dlib.org/dlib/january98/dolin/01dolin.html> [03.05.2003]

Drori & Alon 2003
Drori, O.; Alon, N. (2003). Using document classification for displaying search results. *Journal of information science*. 29(2). 97–106.

Dumais & Chen 2000
Dumais, S.; Chen, H.Hierarchical classification of Web content. *SIGIR 2000: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, Athens, Greece, July 24–28, 2000*. Hrsg.: Belkin, N. J. et al. – New York, NY: ACM Press. 256–263.

Dumais et al. 2002
Dumais, S. T.; Lewis, D. D.; Sebastiani, F. (2002). Report on the workshop on operational text classification systems (OTC-02). *SIGIR forum.* 36(2). WWW, 4p.
<http://www.acm.org/sigir/forum/F2002/sebastiani.pdf> [17.11.2003]

Ellis & Vasconcelos 1999
Ellis, D.; Vasconcelos, A. (1999). Ranganathan and the Net: Using facet analysis to search and organise the World Wide Web. *Aslib proceedings.* 51(1). 3–10.

EPA 2004
Europäisches Patentamt <München, DE> (2004). *2003: Jahresbericht; Annual report; Rapport annuel*. München: EPA. WWW: <http://www.european-patent-office.org/epo/an_rep/index.htm> [29.06.2004]

Factiva 2004
Dow Jones Reuters Business Interactive LLC [Factiva] < New York, NY, US> (2004). *Factiva Research-Produkte*. WWW, 4p.
<http://www.factiva.com/collateral/files/factiva_research_brochure_F-1323-G.pdf> [09.07.2004]

Fall et al. 2003
Fall, C. J.; Törcsvári, A.; Benzineb, K.; Karetka, G. (2003). Automated categorization in the International Patent Classification. *SIGIR forum.* 37(1). WWW, 16p.
<http://www.acm.org/sigir/forum/S2003/CJF_Manuscript_sigir.pdf> [17.11.2003]

Figuerola et al. 2001
Figuerola, C. G.; Zazo Rodríguez, A. F.; Berrocal, J. L. A. (2001). Automatic vs manual categorisation of documents in Spanish. *Journal of documentation*. 57(6). 763–773.

Frank & Paynter 2004
Frank, E.; Paynter, G. W. (2004). Predicting Library of Congress Classifications from Library of Congress Subject Headings. *Journal of the American Society for Information Science and Technology*. 55(3). 214–227.

Fürnkranz 1999
Fürnkranz, J. (1999). Exploiting structural information for text classification on the WWW. *Advances in intelligent data analysis: 3rd international symposium, IDA-99, Amsterdam, NL, August 9–11, 1999; Proceedings.* Hrsg.: Hand, D. H. et al. – Berlin u.a.: Springer. (Lecture notes in computer science; 1642). 487–497.

Gaese 2003
Gaese, V. (2003). Automatische Klassifikation von Presseartikeln in der Gruner + Jahr Dokumentation. *Bibliotheken und Informationseinrichtungen: Aufgaben, Strukturen, Ziele; 29. Arbeits- und Fortbildungstagung der ASpB [...], 8.–11. April 2003 in Stuttgart*. Hrsg.: Brauer, M. – Jülich: ASpB/Sektion 5 im DBV. 401–413.

Garratt et al. 1999
Garratt, A.; Jackson, M.; Burden, P. (1999). Implementing a search engine using an OODB. [Vortrag:] *Conference on Object-oriented Programming, Systems, Languages and Applications (OOPSLA '99), Denver, CO, November 1–5, 1999*. WWW, 17p.
<http://seed.scit.wlv.ac.uk/papers/oopsla/oopsla.html> [29.05.2004]

Gietz 2001
Gietz, P. (2001). *Report on automatic classification systems: For the TERENA activity Portal Coordination.* WWW, 7p. <http://www.daasi.de/reports/Report-automatic-classification.html> [07.02.2003]

Glover et al. 2002
Glover, E. J.; Tsioutsiouliklis, K.; Lawrence, S.; Pennock, D. M. ; Flake, G. W. (2002). Using Web structure for classifying and describing Web pages. [Vortrag:] *WWW 02, Honolulu, USA, May 7–11, 2002*. WWW, 21p. <http://www2002.org/CDROM/refereed/504/index.html> [10.03.2004]

Godby 2001
Godby, J. (2001). Terminology identification from full text: OCLC's WordSmith experience. [Vortrag:] *The Southern Ohio Chapter of the American Society for Information Science & Technology (SO-ASIST) meeting, "Aboutness: Automated indexing & categorization," LexisNexis, Miamisburg, OH, June 21, 2001*. WWW, 31sl.
<http://www.asis.org/Chapters/soasis/events/20010621a.ppt> [24.04.2004]

Godby & Reighart 1998a
Godby, C. J.; Reighart, R. R. (1998). The WordSmith Toolkit. *Annual review of OCLC research 1997*. WWW, 8p.
<http://www.oclc.org/research/publications/arr/1997/godby/godby_wordsmith.htm> [10.04.2003].

Godby & Reighart 1998b
Godby, C. J.; Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewey Decimal Classification. *Proceedings of the 9th ASIS SIG/CR Classification Research Workshop, 25 October 1998, Washington, DC.* – Silver Spring, MD: ASIS. 91–105. Prepr.: WWW, 14p. <http://orc.rsch.oclc.org:5061/papers/sigcr98.html> [16.06.2004]

Godby & Stuler 2001
Godby, C. J.; Stuler, J. (2001). The Library of Congress Classification as a knowledge base for automatic subject categorization. [Vortrag:] *IFLA Preconference "Subject retrieval in a networked environment", Dublin, OH, August 2001*. WWW, 6p. <http://staff.oclc.org/~godby/auto_class/godby-ifla.html> [22.04.2003]. – Ebenso in: *Subject Retrieval in a Networked Environment.* Hrsg.: McIlwaine, I. C. – München: Saur, 2003. (UBCIM publ.; n.s.; 25). 163–169.

Godby & Vizine-Goetz 2000
Godby, J.; Vizine-Goetz, D. (2000). ISKO participants discuss ways librarianship can improve responsiveness of the Web. *OCLC newsletter*. (247). 22–25.

Gödert 2002a
(2002). "Die Welt ist gross - Wir bringen Ordnung in diese Welt". *Information – Wissenschaft und Praxis*. 53(7). 395–400.

Gödert 2002b
Gödert, W. (2002). Potenzial des Einsatzes von Klassifikationen für das Information Retrieval. [Vortrag:] *27. Österreichischer Bibliothekartag, Klagenfurt, 9.-14. September 2002*. 37sl.

Goller et al. 2000
Goller, C.; Löning, J.; Will, T.; Wolff, W. (2000). Automatic document classification: A thorough evaluation of various methods. *Informationskompetenz: Basiskompetenz in der Informationsgesellschaft; Proceedings des 7. Internationalen Symposiums für Informationswissenschaft (ISI 2000)*. Hrsg.: Knorz, G.; Kuhlen, R. – Konstanz: UVK. 145–162.

Gower 2002
Gower, S. (2002). *The Snowfox Relevance Engine: Precision, scalability and a very low cost of ownersip*[sic!]. WWW, 7p. <http://www.snowfox.com/Red.pdf> [23.03.2004]

Greiner et al. 1997
Greiner, R.; Grove, A.; Schuurmans, D. (1997). *On learning hierarchical classifications.* WWW, 11p. <http://www.cs.ualberta.ca/~dale/papers/hier.ps.gz> [23.02.2004]

Hagedorn 2001
Hagedorn, K. (2001). *Extracting value from automated classification tools: The role of manual involvement and controlled vocabularies.* Ann Arbor, MI: Argus Center for Information Architecture. 17p. WWW: <http://argus-acia.com/white_papers/classification.pdf> [10.03.2004]

Haya et al. 2003
Haya, G.; Scholze, F.; Vigen, J. (2003). Developing a Grid-based search and categorization tool. *High energy physics libraries webzine*. (8). WWW, 6p.
<http://library.cern.ch/HEPLW/8/papers/1/> [14.10.2003]

Hees 2004
Hees, H. (2004). *Knowledge Framework*. Mainz: Brainbot Technologies. 27p. WWW:
<http://brainbot.com//site3/dokumente/nextbot_knowledge_framework_whitepaper.pdf> [28.06.2004]

Hickey & Vizine-Goetz 2000
Hickey, T. B.; Vizine-Goetz, D. (2000). The role of classification in CORC. *Annual review of OCLC research 1999*. WWW, 8p.+5fig.
<http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003497> [29.04.2004]

Hiom 1998
Hiom, D. (1998). *Mapping classification schemes.* WWW, V1.0, February 1998. 5p.+12p. app.
<http://www.sosig.ac.uk/desire/class/mapping.html> [03.05.2004]

Hiom 2000
Hiom, D. (2000). SOSIG: An Internet hub for the social sciences, business and law. *Online information review*. 24(1). 54–58.

Hoffmann 2002
Hoffmann, R. (2002). *Entwicklung einer benutzerunterstützten automatisierten Klassifikation von Web-Dokumenten: Untersuchung gegenwärtiger Methoden zur automatisierten Dokumentklassifikation und Implementierung eines Prototyps zum verbesserten Information Retrieval für das xFIND System.* Graz: TU, Dipl.-Arb.

Hobohm 1998
Hobohm, H.-C. (1998). Bibliothekarische Internet-Projekte in Deutschland: Qualität und Nutzerorientierung bei bibliothekarischen Internetprojekten – marketingstrategische Überlegungen zu den neuen Informationsdienstleistungen. [Vortrag*:] Weiter auf dem Weg zur Virtuellen Bibliothek! 3. InetBib-Tagung, 4.-6. März 1998*. WWW, 12p.
<http://eldorado.uni-dortmund.de:8080//bib/98/inetbib3/hobohm.pdf> [01.12.2003]

Insuma 2002
Insuma GmbH <Tübingen, DE> (2002*). Insuma Distributed Search Engine: An Insuma GmbH white paper*. WWW, 11p. <http://www.insuma.de/insuma/download/insuma-whitepaper.pdf> [07.01.2004]

Intology o.J. a
Intology <Canberra, AU> (o.J.) Klarity*: Automated developer tools for text processing*. WWW, 3p. <http://www.intology.com.au/20products/50Klarity/> [24.04.2004]

Intology o.J. b
Intology <Canberra, AU> (o.J.) *Taxonomy Builder*. WWW, 2p.
<http://www.intology.com.au/20products/40taxonomy_Builder/> [24.04.2004]

Inxight 2002c
Inxight Software, Inc. <Sunnyvale, CA, US> (2002). *Inxight*. WWW, 6p.
<http://www.inxight.com/pdfs/inxight_brochure.pdf> [05.06.2004]

Inxight 2004
Inxight Software, Inc. <Sunnyvale, CA, US> (2004). *Inxight SmartDiscovery: Taxonomy and categorization.* WWW, 3p. <http://www.inxight.com/pdfs/Taxonomy_FinalWeb.pdf> [06.05.2004]

Ishida 1998
Ishida, E. (1998). An experiment of automatic classification of books using Nippon decimal classification [Text in japanischer Sprache]. *Library and information science*. (39). 31–45.

Ishikawa 1988
Ishikawa, T. (1988). The man-machine interface aspect of an automatic classification numbering system. *Journal of information processing*. 11(3). 199–205.

Ishikawa et al. 1994
Ishikawa, T.; Nakamura, H.; Nakamura, Y. (1994). UDC number automatic combination system (UDC-AUTCS): Implications for classifying and document[-like object] retrieval. *Knowledge organization and quality management: Proceedings of the 3rd international ISKO conference, Copenhagen, DK, 20–24 June 1994.* Hrsg.: Albrechtsen, H.; Oernager, S. – Frankfurt/Main: Indeks Verl. (Advances in knowledge organization; 4). 328–333.

Jansson o.J.
Jansson, K. (o.J.) *DESIRE: Peer review report* [D3.6a: Automatic classification]. WWW, 7p.
<http://www.desire.org/html/research/deliverables/D3.6/d36peer1.html> [27.04.2004]

Jenkins & Inman 2000
Jenkins, C.; Inman, D. (2000). Adaptive automatic classification on the Web. 11th International *Workshop on Database and Expert Systems Applications, London, Sep. 4–8, 2000; Proceedings*. Hrsg.: Tjoa, A M. et al. – Los Alamitos, CA: IEEE Computer Society. 504–511.

Jenkins et al. 1997
[Jenkins, C.; Jackson, M.; Burden, P.; Wallis, J.] (1997). *Automatic classification of Web resources using Java and Dewey Decimal Classification* [working paper]. WWW, 18p.
<http://www.scit.wlv.ac.uk/~ex1253/classifier/> [12.12.2003]

Jenkins et al. 1998
Jenkins, C.; Jackson, M.; Burden, P.; Wallis, J. (1998). Automatic classification of Web resources using Java and Dewey Decimal Classification. *Computer networks and ISDN systems.* 30(1–7). 646–648.

Jenkins et al. 1999
Jenkins, C.; Jackson, M.; Burden, P.; Wallis, J. (1999). Automatic RDF metadata generation for resource discovery. *Computer networks.* (31). 1305–1320.

Kim & Lee 2002
Kim, J.-H.; Lee, K.-H. (2002). Designing a knowledge base for automatic book classification. *Electronic library.* 20(6). 488–495.

Kleinoeder & Puzicha 2002
Kleinoeder, H. H.; Puzicha, J. (2002). Automatische Kategorisierung am Beispiel einer Pilotanwendung. *Info 7: Information und Dokumentation in Archiven, Mediotheken, Datenbanken.* 17(1). 19–22.

Klinkenberg 1998
Klinkenberg, R. (1998). *Maschinelle Lernverfahren zum adaptiven Informationsfiltern bei sich verändernden Konzepten.* Dortmund: Univ., Dipl.-Arb.
WWW: <http://www-ai.cs.uni-dortmund.de/DOKUMENTE/klinkenberg_98a.pdf> [24.02.2004]

Knox 2003
Knox, T. W. (2003). The secret to eBay success: Is it possible to create a profitable business using the online auctioneer? *Entrepreneur.com,* December 22, 2003. WWW, 4p.
<http://www.entrepreneur.com/article/0,4621,312476,00.html> [09.07.2004]

Koch 1994
Koch, T. (1994). Experiments with automatic classification of WAIS databases and indexing of WWW: Some results from the Nordic WAIS/WWW project. *Internet World & Document Delivery World International 94: Proceedings of the 2nd annual conference, London, May 1994.* Westport, CT: Mecklermedia.112–115.
WWW-Version: 6p. <http://www.lub.lu.se/netlab/documents/nordic_w4.html> [15.04.2004]

Koch 1997a
Koch, T. (1997). Verbesserung der Recherchemöglichkeiten im Internet: Internationaler Überblick. [Vortrag:] *19. Online-Tagung der Deutschen Gesellschaft für Dokumentation: "Die Zukunft der Recherche", INFOBASE, Frankfurt/Main, 15.05.1997.* WWW, 12p.
<http://www.lub.lu.se/tk/demos/DGD97.html> [10.02.2004]

Koch 1997b
Koch, T. (1997). *DESIRE: Development of a European Service for Information on Research and Education: EU project.* Telematics Applications Programme, Telematics For Research. Proposal no: RE 1004 (RE). WWW, 8p. Created: 1995-08-29; L/U: 1997-11-10.
<http://www.ub2.lu.se/desire/desireIindex.html> [01.03.2004]

Koch 1998a
Koch, T. (1998). Mit Robotsoftware und Metadaten zur digitalen Bibliothek: Infrastrukturprojekte und Standardentwicklungen zur Unterstützung der Suche im Internet. [Vortrag:] *MAID München und AKI Stuttgart, 2.–3.3.1998.* WWW, 4p. L/U: 24.02.1998.
<http://www.lub.lu.se/tk/demos/DigBib9803.html> [04.10.2002]

Koch 1998b
Koch, T. (1998). Nutzung von Klassifikationssystemen zur verbesserten Beschreibung, Organisation und Suche von Internetressourcen. *Buch und Bibliothek.* 50(5). 326–335.

Koch 1998c
Koch, T. (1998). *Possible advantages of using traditional library classification systems in Internet services.* WWW, 3p. <http://www.lub.lu.se/tk/demos/mex9808b.html> [15.05.2004]

Koch 1999
Koch, T. (1999). *Automatic classification: DESIRE II workplan WP D3.6a*. WWW, 3p. L/U: 1999-01-05. <http://www.lub.lu.se/tk/desire2/desire2-autoclass-plan.html> [06.02.2004]

Koch 2000
Koch, T. (2000). *Adding automatic classification to a robot-generated subject index*. WWW, 5p. L/U: 2000-05-09. <http://www.lub.lu.se/tk/demos/desire-autoclass.html> [10.04.2003]

Koch & Ardö 2000a
Koch, T.; Ardö, A. (2000). *Automatic classification: DESIRE II D3.6a, Overview of results.* WWW, 8p. L/U: 2000-02-24. <http://www.lub.lu.se/desire/DESIRE36a-overview.html> [15.05.2004]

Koch & Ardö 2000b
Koch, T.; Ardö, A. (2000). *Automatic classification of full-text HTML-documents from one specific subject area: DESIRE II D3.6a, Working Paper 2*. WWW, 27p.+Abb./Tab. L/U: 2000-02-25. <http://www.lub.lu.se/desire/DESIRE36a-WP2.html> [10.04.2003]

Koch & Day 1998
Koch, T.; Day, M. (1998). Review of attempts to apply classification in automated services. Chapter 3 in: *The role of classification schemes in Internet resource description and discovery* (= Specification for resource description methods, part 3).
WWW, 5p. <http://www.lub.lu.se/desire/radar/reports/D3.2.3/f_3.html> [10.04.2003]

Koch & Vizine-Goetz 1999
Koch, T.; Vizine-Goetz, D. (1999). Automatic classification and content navigation support for Web services: DESIRE II cooperates with OCLC. *Annual review of OCLC research 1998.* WWW, 17p. <http://www.oclc.org/research/publications/arr/1998/koch_vizine-goetz/automatic.htm> [10.04.2003].

Koch et al. 1995
Koch, T. ; Ardö, A.; Falcoz, F.; Nielsen, M.; Sandfær, M. (1995). Improving resource discovery and retrieval on the Internet: The Nordic WAIS/World Wide Web Project and the classification of WAIS databases. *Wissen in elektronischen Netzwerken: Strukturierung, Erschliessung und Retrieval von Informationsressourcen im Internet; Eine Auswahl von Vorträgen der 19. Jahrestagung der Gesellschaft für Klassifikation, Basel 1995.* Hrsg.: Hobohm, H. C.; Wätjen, H.-J. – Oldenburg: BIS-Verl. 147–169.

Koch et al. 1996
Koch, T.; Ardö, A.; Brümmer, A. (1996). *The building and maintenance of robot based Internet search services: A review of current indexing and data collection methods.* Prepared to meet the requirements of Work Package 3 of EU Telematics for Research, project DESIRE. Version D3.11v0.3 (Draft version 3). WWW, 116p. L/U: 1996-09-26.
<http://www.lub.lu.se/desire/radar/reports/D3.11/tot.html> [04.05.2003]

Koch et al. 1997
Koch, T.; Day, M.; Brümmer, A.; Hiom, D.; Peereboom, M.; Poulter, A.; Worsfold, E. (1997). *The role of classification schemes in Internet resource description and discovery.* Deliverable no. D3.2 (= Specification for resource description methods, part 3). WWW, 47p.
<http://www.ukoln.ac.uk/metadata/desire/classification/classification.pdf> [05.02.2004]

Koch et al. 1999
Koch, T.; Zettergren, A.-S.; Day, M. (1999). *Provide browsing using classification schemes.* WWW, 14p. <http://www.lub.lu.se/desire/handbook/class.html> [10.04.2003]

Koch et al. 2000
Koch, T.; Ardö, A.; Noodén, L. (2000). *Automatic classification: DESIRE II, deliverable no.* D3.6. WWW, 8p. V1.0, 29 February 2000.
<http://www.desire.org/html/research/deliverables/D3.6/deliverd36.html> [24.04.2004]

Kofax 2004
Kofax Image Products <Irvine, CA, US> (2004). *Moho Classifier*. WWW, 2p.
<http://www.kofax.com/products/mohomine/classifier.asp> [10.03.2004]

Koller & Sahami 1997
Koller, D.; Sahami, M. (1997). Hierarchically classifying documents using very few words. Machine learning: Proceedings of the 14th international conference (ICML '97), Nashville, TN, July 8–12, 1997. Hrsg.: Fisher, D. H. – San Francisco, CA: Morgan Kaufmann. 170–178.

Koster et al. 2001
Koster, C. H. A.; Seutter, M.; Beney, J. (2001). Classifying patent applications with Winnow. *Benelearn 2001: Proceedings of the 11th Belgian-Dutch conference on machine learning, Antwerp University, Belgium, 21 December 2001*. Hrsg.: Hoste, V.; De Pauw, G. – Antwerpen: CNTS – Language Technology Group. 19–26.
WWW: <http://cnts.uia.ac.be/benelearn2001/proceedings/bene01-koster.pdf> [25.03.2004]

Koster et al. 2003
Koster, C. H. A.; Seutter, M.; Beney, J. (2003). Multi-classification of patent applications with Winnow. *Perspectives of system informatics: 5th international Andrei Ershov Memorial Conference (PSI 2003), Akademgorodok, Novosibirsk, Russia, July 9–12, 2003; Revised papers.* Hrsg.: Broy, M. et al. – Berlin u.a.: Springer. (Lecture Notes in Computer Science; 2890). 545–554. – WWW: <http://www.cs.kun.nl/peking/psi2003.pdf> [25.03.2004]

Krellenstein 2001
Krellenstein, M. (2001). Document classification at Northern Light. [Vortrag:] *Search engines and beyond: Developing efficient knowledge management systems; 1999 Search Engine Meeting, Boston, MA, April 19–20 1999.* WWW, 11p. L/U: 14 August 2001.
<http://www.infonortics.com/searchengines/boston99.html> [10.04.2003]

Krier & Zaccà 2002
Krier, M.; Zaccà, F. (2002). Automatic categorisation applications at the European Patent Office. *World patent information.* 24(3). 187–196.

Krüger 1999
Krüger, C. (1999). *Evaluation des WWW-Suchdienstes GERHARD unter besonderer Beachtung der automatischen Indexierung*. Stuttgart: FH Stuttgart – HS für Bibliotheks- und Informationswesen, Dipl.-Arb.

Kuo & Wong 2000
Wong, M.-H.; Kuo, Y.-H. (2000). Web document classification based on hyperlinks and document semantics. *Proceedings of the international workshop on text and Web mining (PRICAI), Melbourne, August 2000*. Hrsg.: Tan, A.-H.; Yu, P. S. – o.O. 44–51.
WWW: <URL http://citeseer.nj.nec.com/kuo00web.html> [24.04.2004]

Kwasnik & Liu 2000
Kwasnik, B. H.; Liu, X. (2000). Classification structures in the changing environment of active commercial websites: The case of eBay.com. *Dynamism and stability in knowledge organization: Proceedings of the 6th international ISKO conference, Toronto, Canada, 10–13 July 2000*. Hrsg.: Beghtol, C. et al. – Würzburg: Ergon. (Advances in knowledge organization; 7). 372–377.

Lamont 2003
Lamont, J. (2003). Dynamic taxonomies: Keeping up with changing content. *KM world.* 12(5). WWW, 7p.
<http://www.kmworld.com/publications/magazine/index.cfm?action=readarticle&Article_ID=1508&Publication_ID=90> [23.03.2004]

Larkey 1998
Larkey, L. S. (1998). Some issues in the automatic classification of U.S. patents. *Learning for Text Categorization: Papers from the 1998 Workshop*. AAAI Press. (Technical Report WS-98-05.) 87–90. – Prepr.: WWW, 4p. <http://citeseer.nj.nec.com/larkey98some.html> [01.01.2004]

Larkey 1999
Larkey, L. S. (1999). A patent search and classification system. *ACM Digital Libraries '99: Proceedings of the 4th ACM conference on digital libraries, Berkeley, CA, Aug. 11–14, 1999.* Hrsg.: Fox, E. A.; Rowe, N. – New York, NY: ACM Press. 79–87. – Prepr.: WWW, 9p. <http://citeseer.nj.nec.com/larkey99patent.html> [01.01.2004]

Larson 1992
Larson, R. R. (1992). Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science*. 43(2). 130–148.

Lauser & Hotho 2003
Lauser, B.; Hotho, A. (2003). Automatic multi-label subject indexing in a multilingual environment. *Research and advanced technology for digital libraries: 7th European conference (ECDL 2003), Trondheim, Norway, August 17–22, 2003; Proceedings*. Hrsg.: Koch, T. et al. – Berlin u.a.: Springer. (Lecture notes in computer science; 2769). 140–151.

Leclercq 1999
Leclercq, I. (1999). INPI, the Internet and electronic commerce. *World patent information.* 21(4). 259–265.

Lepsky 1998
Lepsky, K. (1998). Im Heuhaufen suchen - und finden: Automatische Erschliessung von Internetquellen: Möglichkeiten und Grenzen. *BuB: Forum für Bibliothek und Information.* 50(5). 336–340.

Liddy et al. 1994
Liddy, E. D.; Paik, W.; Woelfel, J. (1994). Use of subject field codes from a machine-readable dictionary for automatic classification of documents. *Proceedings of the 3rd ASIS SIG/CR classification research workshop held at the 55th ASIS annual meeting, Pittsburgh, PA, Oct 25, 1992*. Hrsg.: Fidel, R. et al. – Medford, NJ: Learned Information. (Advances in classification research; 3). 83–100.

Lindholm et al. 2003
Lindholm, J.; Schönthal, T.; Jansson, K. (2003). Experiences of harvesting Web resources in engineering using automatic classification. *Ariadne.* (37). WWW, 6p.
<http://www.ariadne.ac.uk/issue37/lindholm/> [31.10.2003]

Lyon 1999
Lyon M. (1999). Language related problems in the IPC and search systems using natural language. *World patent information.* 21(2). 89–95.

Maggs 1999
Maggs, P. (1999). Planet SOSIG: Exploring social science resources on the Internet. *Ariadne.* (19). WWW, 6p. <http://www.ariadne.ac.uk/issue19/planet-sosig/> [07.01.2004]

Marchiori 1998
Marchiori, M. (1998). The limits of Web metadata, and beyond. *Computer networks and ISDN systems.* 30(1–7). 1–9.

McCallum et al. 1999
McCallum, A.; Nigam, K.; Rennie, J.; Seymore, K. (1999). A machine learning approach to building domain-specific search engines. *16th international joint conference on artificial intelligence (IJCAI-99), Stockholm, July 31 – August 6, 1999. Vol. 2.* San Mateo, CA: Kaufmann. 662–667.

McCallum et al. 2000
McCallum, A. K.; Nigam, K.; Rennie, J.; Seymore, K. (2000). Automating the construction of Internet portals with machine learning. *Information retrieval*. 3(2). 127–163.

McKiernan 1996
McKiernan, G. (1996). Automated categorisation of Web resources: A profile of selected projects, research, products, and services. *New review of information networking*. (2).15–40.

Mitchell 1996
Mitchell, J. S. (1996). The Dewey Decimal Classification at 129: Edition 21 and beyond. *Knowledge organization and change: Proceedings of the 4th International ISKO Conference, Washington, DC, 15-18 July 1996*. Hrsg.: Green, R. – Frankfurt/Main: Indeks Verl.

Mitchell 2001
Mitchell, J. S. (2001). Dewey Decimal Classification: 125 and still growing. *OCLC newsletter*. (254). 27–29.

Mitchell & Vizine-Goetz 2002
Mitchell, J. S.; Vizine-Goetz, D. (2002). *A research agenda for classification*. WWW, 6p. <http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003953> [02.01.2004]

Mladenic 1998
Mladenic, D. (1998). Turning Yahoo into an automatic Web classifier. *ECAI-98: 13th European conference on artificial intelligence, Brighton, UK, 1998.* Hrsg.: Prade, H. – New York, NY: Wiley. 473–474.

Mladenic & Grobelnik 1999
Mladenic, D.; Grobelnik, M. (1999). Assigning keywords to documents using machine learning. *Zbornik radova: Journal of information and organizational sciences.* 23(2). 123–131. [= Proceedings of the 10th international conference on information and intelligent systems (IIS '99), Varazdin, Croatia, 22–24 September, 1999].
Prepr.: WWW, 8p. <http://citeseer.nj.nec.com/mladenic99assigning.html> [26.02.2004]

Möller 1997
Möller, G. (1997). *Relationen in der Universellen Dezimalklassifikation*. WWW, 4p. <http://www.gerhard.de/info/dokumente/dokumentation/gerhard/relationen_udkz.pdf> [06.01.2004]

Möller 2004
Möller-Schwing, G. (2004). Telefongespräch mit dem Verfasser, 05.05.2004.

Möller et al. 1999a
Möller, G.; Carstensen, K.-U.; Diekmann, B.; Wätjen, H. (1999). Automatic classification of the World Wide Web using Universal Decimal Classification. *Online information 99: 23rd International Online Information Meeting, London, 7–9 December 1999*. Hrsg.: McKenna, B. et al. – Oxford, Learned Information Europe. 231–237.

Möller et al. 1999b
Möller, G.; Carstensen, K.-U.; Diekmann, B.; Wätjen, H. (1999). GERHARD: Navigating the Web with the Universal Decimal Classification system. [Vortrag:] *3rd European Conference on Digital Libraries, Paris, 1999*. WWW, 39sl.
<http://www.gerhard.de/info/dokumente/vortraege/ecdl99/ecdl99.ppt> [16.05.2004]

Möller et al. 2000
Möller, G.; Carstensen, K.-U.; Diekmann, B.; Wätjen, H. (2000). Automatic classification of the World-Wide Web using the Universal Decimal Classification. *Classification and information processing at the turn of the millennium: Proceedings of the 23rd annual conference of the Gesellschaft für Klassifikation, Bielefeld, March 10–12, 1999.* Hrsg.: Decker, R.; Gaul, W. – Berlin: Springer. 441–450.
Prepr.: WWW, 12p. <http://www.bis.uni-oldenburg.de/abt1/waetjen/publ/Article.pdf> [01.01.2004]

Moens & Dumortier 1999
Moens, M.-F.; Dumortier, J. (1999). Automatic categorization of magazine articles. *Conferentie informatiewetenschap 1999: Centrum voor Wiskunde en Informatica, 12 november 1999; Proceedings.* Hrsg.: De Bra, P.; Hardman, L.
WWW, 14p. <http://wwwis.win.tue.nl/infwet99/proceedings/moens.html> [03.05.2003]

Moens & Dumortier 2000
Moens, M.-F.; Dumortier, J. (2000). Text categorization: The assignment of subject descriptors to magazine articles. *Information processing and management*. 36(6). 841–861.

Monopoli & Nicholas 2000
Monopoli, M.; Nicholas, D. (2000). A user-centered approach to the evaluation of Subject Based Information Gateways: Case study SOSIG. *Aslib proceedings*. 52(6). 218–231.

Müller 2002
Müller, K. (2002). *Automatische Klassifikation von Textdokumenten*. Hildesheim, Univ., M.-Arb.

Nekrestyanov et al. 1999
Nekrestyanov, I.; O'Meara, T.; Patel, A.; Romanova, E. (1999). Building topic-specific collections with intelligent agents. *Intelligence in services and networks: Paving the way for an open service market; 6th international conference on intelligence and services in networks (IS&N '99), Barcelona, Spain, April 27–29, 1999; Proceedings.* Hrsg.: Zuidweg, H. et al. – Berlin u.a.: Springer. (Lecture notes in computer science; 1597). 70–82.

Neuss & Kent 1995
Neuss, C.; Kent, R. E. (1995). Conceptual analysis of resource meta-information. *Computer networks and ISDN systems.* 27(6). 973–984.
Prepr.: WWW, 18p. <http://www.igd.fhg.de/archive/1995_www95/papers/94/www3.html> [16.05.2004]

Nohr 2003
Nohr, H. (2003). *Grundlagen der automatischen Indexierung: Ein Lehrbuch*. Berlin: Logos.

Northern Light 2003
Northern Light Group <Cambridge, MA, US> (2003). *Northern Light® Enterprise Search Engine: Overview white paper.* WWW, 15p. <http://www.northernlight.com/engine.html> [10.07.2004]

OCLC 2001
Anon. (2001). News from OCLC: c) OCLC Cooperative Online Resource Catalog. *Program: Electronic library and information systems.* 35(1).

OCLC 2003
OCLC Online Computer Library Center, Inc. <Dublin, OH, US> (2003). *Automatic classification research at OCLC*. WWW, 4p.
<http://www.oclc.org/research/projects/auto_class/default.htm> [20.11.2003].

OCLC o.J. a
OCLC Online Computer Library Center, Inc. <Dublin, OH, US> (o.J.) *Scorpion database design*. WWW, 3p. [vermutlich nach 2001].
<http://www.oclc.org/research/software/scorpion/documentation/designdatabase.htm> [23.02.2004]

OCLC o.J. b
OCLC Online Computer Library Center, Inc. <Dublin, OH, US> (o.J.) *Connexion: Integrated cataloging service.* WWW, 2p.
<http://www.oclc.org/connexion/default.htm> [25.03.2004].

OCLC o.J. c
OCLC Online Computer Library Center, Inc. <Dublin, OH, US> (o.J.) *WebDewey user guide*. WWW, 56p. <http://www.oclc.org/support/documentation/dewey/webdewey_userguide/> [25.03.2004].

OFFIS 1998
OFFIS <Oldenburg, DE> (1998). *Jahresbericht – 1997.* WWW, Mai 1998. [darin: 9.3. Ein systematisches Verzeichnis des deutschen WWW. 5p.]
<http://www.offis.uni-oldenburg.de/publikationen/jahresbericht/jb1997/p9_3.php> [10.03.2004]

Oh et al. 2000
Oh, H. J.; Myaeng, S. H.; Lee, M. H. (2000). A practical hypertext categorization method using links and incrementally available class information. *SIGIR 2000: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, Athens, Greece, July 24–28, 2000*. Hrsg.: Belkin; N. J. et al. – New York, NY: ACM Press. 264–271.

Patel 2002
Patel, D. (2002). Organizing the Web: A faceted approach. [Vortrag:] *Workshop on information resource management, DRTC, Bangalore, 13–15 March 2002*. WWW, 10p.
<https://drtc.isibang.ac.in/retrieve/200/Paper-CG.PDF> [18.11.2003]

Pejtersen 1993
Pejtersen, A. M. (1993). A new approach to design of document retrieval and indexing systems for OPAC users. *Online information '93: 17th international online information meeting proceeding; London, 7–9 December 1993*. Hrsg.: Raitt, D. I.; Jeapes, B. – Medford, NJ: Information Today. 273–290.

Pierre 2000
Pierre, J. M. (2000). Practical issues for automated categorization of Web sites. [Vortrag:] *ECDL 2000 workshop on the semantic Web, 21 September 2000, Lisbon Portugal*. WWW, 9p.
<http://www.ics.forth.gr/isl/SemWeb/proceedings/session3-3/html_version/semanticweb.html> [03.05.2003]

Pierre 2001
Pierre, J. M. (2001). On the automated classification of Web sites. *Linköping electronic articles in computer and information science*. 6(1). WWW, 15p. <http://www.ep.liu.se/ea/cis/2001/001/> [17.02.2004]

Porter 1980
Porter, M. F. (1980). An algorithm for suffix stripping. *Program.* 14(3). 130–137. [zit. n. Koch & Ardö 2000b]

Prabowo et al. 2002a
Prabowo, R.; Jackson, M.; Burden, P.; Knoell, H.-D. (2002). Ontology-based automatic classification for the Web pages: Design, implementation and evaluation. *WISE '00: Proceedings of the 3rd International Conference on Web Information Systems Engineering, Singapore, 12–14 Dec. 2002*. Hrsg.: Ling, T. W. et al. – Los Alamitos, CA: IEEE Computer Society. 182–191.

Prabowo et al. 2002b
Prabowo, R.; Jackson, M.; Burden, P.; Knöll, H.-D. (2002). Ontology-based automatic classifier for classifying the Web pages. *Proceedings of ETCE2002: ASME Engineering Technology Conference on Energy, Houston, TX, 4–5 February 2002*. o.O. 12p.

Projekt GERHARD2 2001
Anon. (2001). *Projekt: GERHARD2: German Harvest Automated Retrieval and Directory II.* WWW, 3p. <http://www-is.informatik.uni-oldenburg.de/forschung/1657.html> [25.01.2004]

Projektantrag GERHARD II o.J.
Anon. (o.J.) *Projektantrag GERHARD II*. WWW, 8p. <http://www.gerhard.de/info/gerhard2.html> [08.02.2003]

Quaresma & Rodrigues 2002
Quaresma, P.; Rodrigues, I. P. (2002). Automatic classification and intelligent clustering for WWWeb information retrieval systems. *Journal of information, law and technology.* (2). WWW, 12p. <http://elj.warwick.ac.uk/jilt/00-2/quaresma.html> [10.03.2004]

Quint 1999
Quint, B. (1999). Company dossier product emerges from LEXIS-NEXIS' SmartIndexing technology. *NewsBreaks, Information Today Inc.* (November 8). WWW, 3p. <http://infotoday.mondosearch.com/> [10.04.2003]

Rapke 2001
Rapke, K. (2001). Automatische Indexierung von Volltexten für die Gruner + Jahr Pressedatenbank. *nfd Information – Wissenschaft und Praxis.* 52(5). 251–262.

Rapoza 2002
Rapoza, J. (2002). Three paths to sorting content. *eWeek enterprise news & reviews.* (July 15). WWW, 3p. <http://www.eweek.com/print_article/0,1761,a=29100,00.asp> [13.07.2004]

Ravid 2002
Ravid, Y. (2002). *GammaWare Technology*. GammaSite, Inc., June 2002. WWW, 43sl. <http://www.univ-orleans.fr/SCIENCES/LIFO/Manifestations/CAp2002/enligne/GammaSite-CAp2002.ppt> [15.12.2003]

Recommind 2004
Recommind Inc. <Berkeley, CA, US> (2004). *MindServer Categorization: Product specifications*. WWW, 2p. <http://www.recommind.com/categorization.html> [06.04.2004]

Recommind o.J.
Recommind Inc. <Berkeley, CA, US> (o.J.) *MindServer Technology white paper*. Berkeley, CA: Recommind. 12p.

Renz 2001
Renz, M. (2001). Automatische Inhaltserschliessung im Zeichen von Wissensmanagement. *nfd Information – Wissenschaft und Praxis*. 52(2). 69–78.

Robbins 1999
Robbins, F. (1999). An exploration of the application of classification systems as a method of resource delivery on the World Wide Web. *Cataloguing Australia.* 25(1/4). 60–65.

Sachse et al. 1998
Sachse, E.; Liebig, M.; Gödert, W. (1998). *Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt*. Köln: FH Köln, Fachbereich Bibliotheks- und Informationswesen. 65p. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; 14).

Salton & McGill 1987
Salton, G.; McGill, M. J. (1987). *Information Retrieval: Grundlegendes für Informationswissenschaftler.* Hamburg usw.: McGraw-Hill.

Schmeer & Sidlo 1998
Schmeer, D.; Sidlo, C. (1998). Automatic topical indexing at LEXIS-NEXIS. *19th annual national online meeting, proceedings, New York, NY, May 12–14, 1998.* Hrsg.: Williams, M. E. – Medford, NJ: Information Today. 337–345.

Schönthal 2003
[Schönthal, T.] (2003). *Engine-e behind the screen: http://engine-e.lub.lu.se/.* WWW, 3p. <http://www.lub.lu.se/knowtech/projects/engine-e/behind-the-screen.html> [27.05.2004]

Scholze 2003
Scholze, F. (2003). GRACE: Eine Grid-basierte und kategoriebildende Suchmaschine. *BIT online.* 6(2). 155–159.

Schultz o.J.
Schultz, L. (o.J.). *Hierarchies and eBay*. WWW, 25p. [vermutlich ca. 2002] <http://www.tarleton.edu/~schultz/ebayfinalpaper1.doc> [23.03.2004]

Sebastiani 2001
Sebastiani, F. (2001). Organizing and using digital libraries by automated text categorization. *Atti del Workshop su Intelligenza artificiale per i beni culturali e le biblioteche digitali, Bari, 25 settembre 2001; Proceedings of the AI*IA workshop on artificial intelligence for the cultural heritage and digital libraries*. Hrsg.: Bordoni, L; Semeraro, G. – o.O. 93–94.
Prepr.: WWW, 3p. <http://faure.iei.pi.cnr.it/~fabrizio/Publications/AIIA01.pdf> [06.01.2004]

Sebastiani 2002a
Sebastiani, F. (2002). Automated text categorization: Tools, techniques and applications. [Vortrag:] *Centre National de Recherche Technologique, Text Indexing Seminar, Rennes, France, April 3, 2002.* WWW, 30p. <http://tim.irisa.fr/veille/text-mining/fabrizi.pdf> [06.01.2004]

Sebastiani 2002b
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys.* 34(1). 1–47.

Sebastiani 2004
Sebastiani, F. (2004). Text categorization. *Text mining and its applications.* Hrsg.: Zanasi, A. – Southampton, UK, WIT Press. (Management information systems; 3). [im Druck]
Prepr.: WWW, 23p. <http://faure.iei.pi.cnr.it/~fabrizio/Publications/TM03.pdf> [06.01.2004]

Shafer 1996
Shafer, K. (1996). *A brief introduction to Scorpion.* WWW, 2p. <http://orc.rsch.oclc.org:6109/bintro.html> [10.04.2003]

Shafer 1997a
Shafer, K. E. (1997). Automatic subject assignment via the Scorpion system. *Annual review of OCLC research 1996.* WWW, 2p. <http://www.oclc.org/research/publications/arr/1996/> [10.04.2003]

Shafer 1997b
Shafer, K. (1997). Scorpion helps catalog the Web. *Bulletin of the American Society for Information Science.* 24(1). 28–29.

Shafer 1997c
Shafer, K. (1997). Scorpion and automatic subject assignment for Web search engines. [Vortrag:] *AMIGOS 4, 06.11.1997*. WWW, 21sl. <http://orc.rsch.oclc.org:6109/amigos97/> [01.03.2004]

Shafer 1998
Shafer, K. E. (1998). Evaluating Scorpion results. *Annual review of OCLC research 1997*. WWW, 6p. <http://www.oclc.org/research/publications/arr/1997/shafer/eval_scorpion/eval_sc.html> [10.04.2003].

Shafer & Thompson 1997
Shafer, K.; Thompson, R. (1997). *Scorpion: SMART weighting schemes*. WWW, 2p. L/U: Mar 21, 1997. <http://orc.rsch.oclc.org:6109/smart_weight.html> [18.02.2004]

Shafer et al. 1997
Shafer, K.; Thompson, R.; Tkac, V. (1997). *Scorpion: Dewey database design*. WWW, 9p. L/U: Nov 4, 1997. <http://orc.rsch.oclc.org:6109/dewey_db_design.html> [10.04.2003]

Shafer et al. 1999
Shafer, K.; Subramanian, S.; Fausey, J. (1999). *Measures for evaluating automatic subject assignment of electronic resources.* WWW, 12p. <http://orc.rsch.oclc.org:6109/measures.html> [01.03.2004]

Shewhart 2001a
Shewhart, M. (2001). Portable classification tools. *National online 2001: Proceedings*. New York, NY: Information Today. 441–449.

Shewhart 2001b
Shewhart, M. (2001). Portable classification tools. [Vortrag:] *Aboutness: Automated indexing & categorization, Southern Ohio Chapter of the American Society for Information Science & Technology, Dayton, OH, June 21, 2001*. WWW, 40sl.
<http://www.asis.org/Chapters/soasis/events/20010621c.ppt> [22.02.2004]

Shimizu 2003
Shimizu, M. (2003). The OWAKE automatic IPC classification tool at IPCC. [Vortrag:] *2nd EPIDOS users' meeting, Vienna, 23–24 October 2003.* WWW, 20sl. <http://www.european-patent-office.org/epidos/conf/jpinfo/2003/pdf/shimizu_masakazu_ipcc.pdf> [16.05.2004]

Sieben-D 2003
7d Software GmbH & Co. KG <Hamburg, DE> (2003). *Whitepaper: Technologiebeschreibung, September 2003*. WWW, 21p. <http://www.7d-ag.de/pdf/Whitepaper-7d-D-03.pdf> [14.06.2004]

Smith 2002
Smith, H. (2002). Automation of patent classification. *World patent information*. 24(4). 269–271.

Srinivasan o.J.
Srinivasan, P. (o.J.) *DESIRE: Peer review report* [D3.6a: Automatic classification]. WWW, 6p. <http://www.desire.org/html/research/deliverables/D3.6/d36peer2.html> [27.04.2004]

Srishaila 2001
Srishaila, S. (2001). Tools for assigning subjects to e-documents: A step towards organizing Internet resources. [Vortrag:] *Workshop on Multimedia and Internet Technologies, DRTC, Bangalore, 26–28 Feb. 2001*. WWW, 18p. <https://drtc.isibang.ac.in/retrieve/258/Smitha.pdf> [18.11.2003]

Subramanian & Shafer 1998
Subramanian, S.; Shafer, K. E. (1998). Clustering. *Annual review of OCLC research 1997*. WWW, 6p. <http://www.oclc.org/research/publications/arr/1997/shafer/clustering/clustering.htm> [10.04.2003].

Sykes 2001
Sykes, J. (2001). *The value of indexing: A white paper prepared for Factiva, a Dow Jones and Reuters company*. WWW, 9p. <http://www.factiva.com/infopro/indexingwhitepaper.pdf> [07/04/2003]

Sykes 2003
Sykes, J. (2003). *Making solid business decisions through Intelligent Indexing taxonomies: A white paper prepared for Factiva, a Dow Jones and Reuters company*. WWW, 8p. <http://www.factiva.com/collateral/download_brch.asp?node=menuElem1506#white> [09/07/2004]

Thompson et al. 1997
Thompson, R.; Vizine-Goetz, D.; Shafer, K. (1997). Evaluating Dewey concepts as a knowledge base for automatic subject assignment. *ACM Digital Libraries '97: Proceedings of the 2nd ACM international conference on digital libraries, Philadelphia, PA, July 23–26, 1997*. Hrsg.: Allen, R. B. – New York, NY: ACM Press. 37–46.

Thunderstone o.J.
Thunderstone Software, Expansion Programs International, Inc. <Cleveland, OH> (o.J.). *Thunderstone Texis Categorizer*. WWW, 2p. <http://www.thunderstone.com/texis/site/pages/Categorizer.html> [13.07.2004]

Topic clustering 2002
Anon. (2002). *Topic clustering*. WWW, 4p. <http://www.faganfinder.com/search/clustering.shtml> [12.12.2002]

Tóth 2002
Tóth, E. (2002). Innovative solutions in automatic classification: A brief summary. *Libri: International journal of libraries and information services.* 52(1). 48–53.

Tröger 1998
Tröger, B. (1998). Und wie halten Sie es mit der Internet-Erschliessung? Bibliothekarische Gretchenfragen von IBIS bis GERHARD. *Bibliotheksdienst.* 32(11). 1922–1930.

Verity 2001
Verity Deutschland GmbH <Grossostheim, DE> (2001). *Verity Intelligent Classifier: Organisieren Sie Ihre Informationen entsprechend Ihren Unternehmensanforderungen*. WWW, 2p. <http://www.verity.com/de/de_pdf/MK0340_ICData_de.pdf> [07.02.2003]

Vizine-Goetz 1996
Vizine-Goetz, D. (1996). Online classification: Implications for classifying and document[-like object] retrieval. *Knowledge organization and change: Proceedings of the 4th International ISKO Conference, Washington, DC, 15–18 July 1996.* Hrsg.: Green, R. – Frankfurt/Main: Indeks Verl. – Prepr.: WWW, 6p. <http://orc.rsch.oclc.org:6109/dvgisko.htm> [01.03.2004]

Vizine-Goetz 1997a
Vizine-Goetz, D. (1997). From book classification to knowledge organization: Improving Internet resource description and discovery. *Bulletin of the American Society for Information Science*. 24(1). 24–27.

Vizine-Goetz 1997b
Vizine-Goetz, D. (1997). Classification research at OCLC. *Annual review of OCLC research 1996*. WWW, 8p. <http://digitalarchive.oclc.org/da/ViewObject.jsp?fileid=0000003512:000000091448&reqid=5520> [24.07.2004]

Vizine-Goetz 1998a
Vizine-Goetz, D. (1998). Subject headings for everyone: Popular Library of Congress Subject Headings with Dewey numbers. *OCLC newsletter*. (233).29–33.

Vizine-Goetz 1998b
Vizine-Goetz, D. (1998). Popular LCSH with Dewey numbers: Subject Headings for Everyone. *Annual review of OCLC research 1997*. WWW, 9p. <http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003449> [01.03.2004].

Vizine-Goetz 1999a
Vizine-Goetz, D. (1999). NetLab / OCLC collaboration seeks to improve Web searching. *OCLC newsletter*. (240).30–36.

Vizine-Goetz 1999b
Vizine-Goetz, D. (1999). Using library classification schemes for Internet resources. [Vortrag:] *OCLC Internet Cataloging Project Colloquium, San Antonio, TX, 19 January 1996.* WWW, 8p. L/U: 11/23/1999. <http://staff.oclc.org/~vizine/Intercat/vizine-goetz.htm> [23.02.2004]

Vizine-Goetz 2001
Vizine-Goetz, D. (2001). Dewey research: New uses for the DDC. *OCLC newsletter*. (254). 24–26.

Vizine-Goetz 2002
Vizine-Goetz, D. (2002). Classification schemes for Internet resources revisited. *Journal of Internet cataloging*. 5(4). 5–18.

Vizine-Goetz et al. 2000
Vizine-Goetz, D.; Ardö, A.; Godby, J.; Houghton, A.; Koch, T.; Reighart, R.; Thompson, R. (2000). Browsing engineering resources on the Web: A general knowledge organization scheme (Dewey) vs. a special scheme (EI). [Vortrag:] *6th International ISKO Conference, Toronto, Canada, 10-13 July 2000.* WWW, 6p.
<http://www.lub.lu.se/tk/publ/OCLC_NetLab_ISKO6.html> [10.04.2003]

Voss & Gutenschwager 2001
Voss, S.; Gutenschwager, K. (2001). *Informationsmanagement.* Berlin: Springer.

Wätjen 1998a
Wätjen, H.-J. (1998). Automatisches Sammeln, Klassifizieren und Indexieren von wissenschaftlich relevanten Informationsressourcen im deutschen World Wide Web: Das DFG-Projekt GERHARD. *Host-Retrieval und Global Research: 20. Online-Tagung der DGD, 5.–7. Mai 1998.* Hrsg.: Ockenfeld, M. – Frankfurt/Main. – Prepr.: WWW, 5p.
<http://www.gerhard.de/info/dokumente/vortraege/DGD_1998/DGD-Vortrag.htm> [01.12.2003]

Wätjen 1998b
Wätjen, H.-J. (1998). GERHARD: Automatisches Sammeln, Klassifizieren und Indexieren von wissenschaftlich relevanten Informationsressourcen im deutschen World Wide Web. *BIT Online*. 1(4). 279–291.

Wätjen et al. 1998
Wätjen, H.-J.; Diekmann, B.; Möller, G.; Carstensen, K.-U. (1998). *Bericht zum DFG-Projekt: GERHARD: German Harvest Automated Retrieval and Directory, http://www.gerhard.de. Stand: 16.6.1998.* Oldenburg, BIS der Univ. – WWW, 34p.
<http://www.gerhard.de/info/dokumente/dokumentation/gerhard/bericht.pdf> [06.01.2004]

Wallis & Burden 1995
Wallis, J.; Burden, P. (1995). Towards a classification-based approach to resource discovery on the Web. [Vortrag:] *4th International W4G Workshop on Design and Electronic Publishing, Abingdon, UK, 20–22 Nov. 1995*. WWW, 6p. <http://www.scit.wlv.ac.uk/wwlib/position.html> [18.02.2004]

Walther 2001
Walther, R. (2001). *Möglichkeiten und Grenzen automatischer Klassifikationen von Web-Dokumenten.* Bern: Univ., Liz.-Arb.

Ward 1999
Ward, J. (1999). Indexing and classification at Northern Light. [Vortrag:] *CENDI Subject Analysis and Retrieval Working Group conference "Controlled vocabulary and the Internet", September 29, 1999*. WWW, 29sl. <http://www.dtic.mil/cendi/presentations/ward.ppt> [01.03.2004]

Wasson 2001
Wasson, M. (2001). Classification technology at LexisNexis. [Vortrag:] *SIGIR 2001 workshop on operational text classification, New Orleans, LA, September 13, 2001*. WWW, 4p.+15sl. <http://www.daviddlewis.com/events/otc2001/presentations/otc01-wasson-paper.txt> [18.02.2004], <http://www.daviddlewis.com/events/otc2001/presentations/otc01-wasson-slides.ppt> [18.02.2004]

Wheatley 2000
Wheatley, A. (2000). Subject trees on the Internet: A new rôle for bibliographic classification? *Journal of Internet cataloging*. 2(3/4). 115–141.

Winkler 1997
Winkler, R. (1997). *Automatisches Klassifizieren von Zeitungsartikeln*. München: TU, Dipl.-Arb.

Woodward 1996
Woodward, J. (1996). Cataloging and classifying information on the Internet. *Annual review of information science and technology, vol. 31*. Hrsg.: Williams, M. E. – Medford, NJ: Information Today. 189–220.

Worsfold 1997
Worsfold, E. (1997). Planet SOSIG: A new Internet role for Europe's librarians. *Ariadne*. (9). WWW, 7p. <http://www.ariadne.ac.uk/issue9/planet-sosig/> [07.01.2004]

Yang 1999
Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*. 1(1/2). 69–90.

Yang & Liu 1999
Yang, Y. & Liu, X. (1999). A re-examination of text categorization models. *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, CA, 1999*. New York, NY: ACM Press. 42–49.

Zamir & Etzioni 1998
Zamir, O.; Etzioni, O. (1998). *Grouper: A dynamic clustering interface to Web search results*. WWW, 16p. <http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html> [08.04.2003]

Zettergren 2000
Zettergren, A.-S. (2000). *DESIRE II: Development of a European Service for Information on Research and Education: A continuation of the DESIRE project*. Telematics Applications Programme, Telematics For Research. Proposal no.: RE 4004 (RE). WWW, 6p. L/U: 2000-03-01. <http://www.lub.lu.se/desire/desireIIindex.html> [10.04.2003]

Zins 2002
Zins, C. (2002). Models for classifying Internet resources. *Knowledge organization*. 29(1). 20–28.

Eidesstattliche Erklärung

Hiemit versichere ich, die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben.

Wien, den 27. Juli 2004

_____

Dr. Otto Oberhauser