VZG-Projekt Colibri

Überblick, Stand, Ergebnisse
Juli-Dezember 2003

Ulrike Reiner

VZG-Colibri-Bericht 1/2003

Verbundzentrale des
Gemeinsamen Bibliotheksverbundes (VZG)
Platz der Göttinger Sieben 1
D-37073 Göttingen
http://www.gbv.de/

Name des Dokuments

colibri01.doc (Colibri-Aufnahme aus : http://www.interlog.com/~barrow/mango.jpg)

Inhalt

Versionen

colibri01-03-07-18.doc	Entwurf
colibri01-03-10-02.doc	Kleinere Korrekturen bzw. Modifikationen
colibri01-03-11-26.doc	Kapitel 4 (Fortsetzung)
colibri01-03-12-31.doc	Kapitel 3 und 4 (Überarbeitung)
colibri01-04-03-11.doc	Minimale Modifikationen (Einarbeitung von Anregungen)

1. Einleitung

"Hummingbirds can fly
right, left, up, down,
backwards,
even upside down."

(http://portalproductions.com/h/behavior.htm)

Name und Ursprung des im Frühling 2003 aufgenommenen VZG-Projektes „Colibri" gehen auf das Pica-Projekt[1] „Colibri (COntext generation and LInguistic tools for Bibliographic Retrieval Interfaces)" zurück. Colibri erfährt im Rahmen des DDB-Projektes „DDC Deutsch / WebDewey Deutsch" in der VZG eine Renaissance. Intendiert ist eine Verbesserung / Erweiterung bestehender inhaltlicher Erschliessung und Suchmöglichkeiten (um z.B. eine systematische, themenkonzentrierte, fachgebietsbezogene Suche) unter Verwendung der Dewey-Dezimalklassifikation „DDC" (Dewey Decimal Classification) - ein in mehr als 135 Ländern und in über 30 Sprachen verbreitetes System zur Wissensorganisation[2].

Ein VZG-Projektziel innerhalb „Colibri" ist die automatische Klassifikation von (GVK-PLUS[3]-) Titeln in eine der obersten 1000 Klassen der DDC-Hierarchie, d.h. automatische Vergabe einer DDC-Notation zu einem (GVK-PLUS)-Titeldatensatz. Ein (Zeitschriften-) Titel, der von „Kolibris" (lat. „Trochili" bzw. engl. „hummingbirds") handelt, sollte demnach die Notation „598" erhalten. Dieser Ansatz liefert eine noch gröbere Klassifikation als die Basisklassifikation mit ihren ca. 2.100 Klassen[4]. Innerhalb des VZG-Projektes „Colibri" wird folgende offene Frage untersucht:

Q1-COLIBRI:
„Ist es möglich, mit den im GBV (oder aus anderen Quellen erhältlichen) verfügbaren Daten eine inhaltlich stimmige Titel-Klassfikation aller GVK-PLUS-Titel (in die ersten 1000 Klassen) in akzeptabler (Computer-)Laufzeit automatisch zu erzielen?"

Zur Beantwortung der Frage ist Forschungs- und Entwicklungsaufwand nötig. Dieser Bericht stellt den momentanen Stand des VZG-Projektes „Colibri" dar. Nach einleitenden Worten zur automatischen Klassifikation mit Hinweis auf laufende Forschungsprojekte wird die erstellte prototypische Test- und Entwicklungsumgebung beschrieben und die ersten Testergebnisse werden vorgestellt, diskutiert und bewertet.

[1] Pica, Jaarsverlag Annual report 97, ISSN 0168-992, ISBN 90-70311-87-9, S. 18

[2] Dewey Decimal Classification and Relative Index. Ed. 21 (devised by M. Dewey; ed. by J.S. Mitchell, J. Beall, W.E. Matthews, G.R. New), Vol. 1, Forest Press, OCLC Online Computer Library Center, Inc., Albany, New York, 1996, p. xxxi.

[3] Gemeinsamer VerbundKatalog mit Online Contents (http://www.gbv.de/du/dbasesinfo/gvk-plus.shtml)

[4] http://www.gbv.de/du/sacher/kap1.shtml

2. Automatische Klassifikation

1968 schreibt Gerard Salton, ein Pionier des Information Retrieval, in seinem als Klassiker geltenden Werk[5]:

„Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information. In recent years, this whole subject has received an increasing amount of attention ... That there are substantial problems in the information field, everyone is agreed upon: More and more information is generated and put into circulation; the existing tools, classification schedules, and storage arrangements are often inadequate, ...; and it generally becomes more difficult and more expensive to get to know what one needs to know. ... important theoretical problems are unsolved: What exactly is the content or meaning of a document? What are the linguistic devices used to carry meaning? To what extent can individual words, or word groups, in a text be said to carry and maintain a well-defined, controlled meaning? How can one isolate the content-bearing units if they exist? And so on."

Nach über 30 Jahren haben diese Aussagen ihre Aktualität nicht eingebüßt, sondern – wie bekannt – geht das Verbreiten von elektronischer Information seit Inbetriebnahme des ersten Webservers „info.cern.ch"[6] immer rasanter vonstatten. Diese Entwicklung macht es nötig, Verfahren zur Speicherung und Wiedergewinnung von Information weitergehend zu automatisieren. Da die Titelaufnahmen natürlichsprachliche Anteile enthalten, ein schwieriges, ev. nur teilweise lösbares oder unlösbares Problem[7].

Probleme der natürlichen Sprache treten bei der Beantwortung von Anfragen (nach Fakten) an ein Datenbanksystem wie bspw. „Welche Flugrichtungen beherrscht ein Kolibri?" üblicherweise nicht auf, während sie charakteristisch für das Gebiet des Information Retrieval sind. Um eine Anfrage: „Zu welcher Klasse gehören Dokumente (Titel), die von Kolibris handeln?" beantworten zu können, muss die Bedeutung der Anfrage verstanden werden. Da sich im Unterschied zu künstlichen Sprachen Syntax und Semantik natürlicher Sprachen i. a. nicht voneinander trennen (d.h. inhaltliche Aspekte der natürlichen Sprache nicht formalisieren) lassen, sind heuristische Methoden notwendig.

In der Literatur sind unterschiedliche Verfahren zur Inhaltsanalyse und automatischen Klassifikation vorgeschlagen worden, dazu zählen u.a. statistische und linguistische Methoden und Verfahren aus den Gebieten des Information Retrieval[8] und der Künstlichen Intelligenz mit Teilgebieten wie „Content-Based Retrieval", „Textmining" „Produktionssysteme", „maschinelles Lernen", „Data Mining"[9] und „Neuronale Netze". Üblich sind Ermittlung von

[5] Gerard Salton: Automatic Information Organization and Retrieval. McGraw-Hill, Inc., New York u.a., 1968, pp. v; 1; 3

[6] http://www.w3.org/People/Berners-Lee/ShortHistory.html

[7] "Although research has been ongoing for over two decades now there is no sign that automatic procedures are sufficiently developed to replace manual classifiction." aus: Rita Marcella; Robert Newton: A New Manual of Classification. Gower Publ., Hampshire (England) & Vermont (USA), 1994, p. 270

[8] Reginald Ferber: Information Retrieval – Suchmodelle und Data-Mining - Verfahren für Text-sammlungen und das Web. dpunkt.verlag, Heidelberg, 2003

[9] Usama M. Fayyad; Gregory Piatetsky-Shapiro; Padhraic Smyth; Ramasamy Uthurusamy (eds.): Advances in Knowledge Discovery and Data Mining. AAAI Press u.a., 1996

inhaltstragenden Wörtern, Diskriminatoren und Banalwörtern, Bestimmung von Häufigkeiten und Ähnlichkeiten (und deren Verteilungen), Aufbau von (Wortstamm-, Synonym-, Phrasen-, Begriffs) Wörterbüchern, Thesauri und Wissensbasen (bislang weniger üblich). Die Aufgabe des (automatischen) Klassifizierens besteht darin, eine Menge in Teilmengen so aufzuteilen, dass die individuellen Eigenschaften der Elemente innerhalb einer Teilmenge ignoriert werden und die gebildeten Teilmengen sich aufgrund des Klassifizierungskriteriums voneinander unterscheiden lassen. Teilaufgaben sind Extraktion von Klassencharakeristika (1), Klassengenerierung (2) und Objektklassifikation (3).

Eine intellektuell vorgenommene Realisierung der Teilaufgabe (2) liegt bspw. von Melvil Dewey vor, der 1873 eine Klassifikation des menschlichen Wissens konzipiert und diese 1876 publiziert hat. Die nach ihm benannte DDC (Dewey Decimal Classification) wird in der Abt. Dezimal-Klassifikation der „Library of Congress" (LoC) beständig weiterentwickelt, gepflegt und verwendet[10]. Seit Juli 2003 ist die 22. DDC-Auflage in WebDewey verfügbar.[11] DDC ist das wichtigste bibliographische Klassifikationssystem und gehört zusammen mit der UDC (Universal Decimal Classification) und LCC (Library of Congress Classification) zu den drei größten Klassifikationssystemen[12]. DDC ist ein hierarchisches Klassfikationssystem. Auf jeder Hierarchie-Ebene werden jeweils zehn Wissensbereiche in jeweils zehn speziellere Bereiche für die nächste Hierarchie-Ebene aufgeteilt, wobei zwischen Klasse und Teilklasse eine Abstraktionsbeziehung besteht. Klassen werden abkürzend durch (ev. mit Markierungszeichen angereicherte) Ziffernfolgen notiert, aus denen die Stellung in der Hierarchie gut ablesbar ist und dadurch ausserdem eine sprachunabhängige Notation zur Verfügung stellt („160" steht für „Logik", „logics", „logica", „logique", „lógica", ...) :

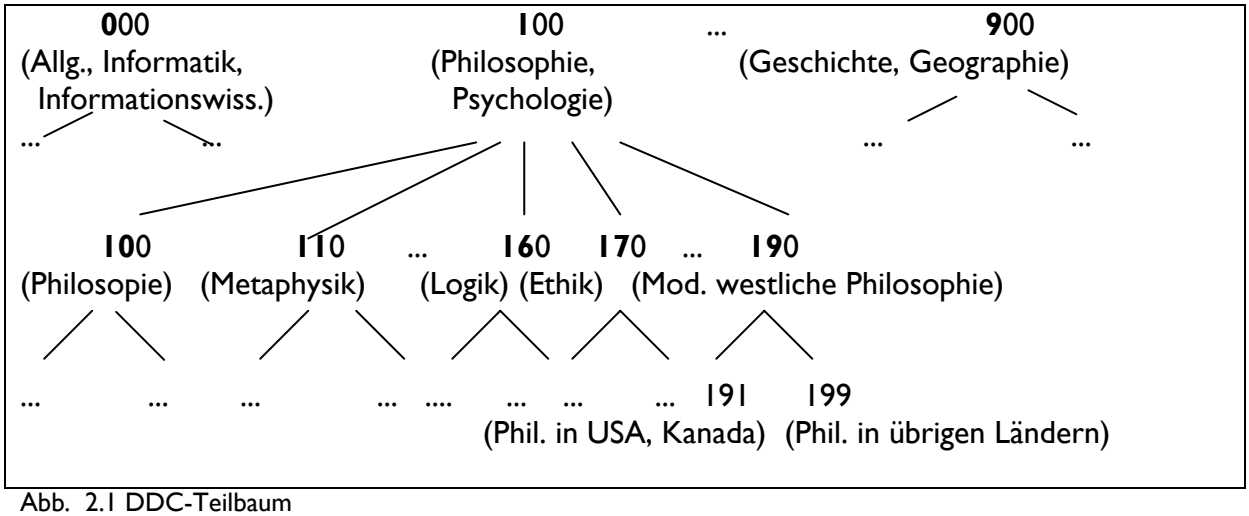


Abb. 2.1 DDC-Teilbaum

[10] Dewey Decimal Classification and Relative Index. Ed. 21 (devised by M. Dewey; ed. by J.S. Mitchell, J. Beall, W.E. Matthews, G.R. New), Vol. 1, Forest Press, OCLC Online Computer Library Center, Inc., Albany, New York, 1996, p. xxxi.

[11] http://www.oclc.org/dewey/products/webdewey/about.htm

[12] Jennifer Rowley: Organizing Knowledge - An Introduction to Information Retrieval. 2nd ed. Gower Publ., Hants (England) & Vermont (USA), 1995, Chap. 14: Bibliographic classification schemes, pp. 200

In Abbildung 2.1 ist ein Teil des DDC-Baumes aus den ersten drei Hierarchie-Ebenen skizziert. Die obersten 10 Klassen (**0**00 – **9**00) werden „tens“ bzw. „(main) classes“ genannt, die der nächsten Ebene (**00**0-**99**0) „hundreds“ bzw. „divisions“ und die Klassen der nächsten Ebene (**000-999**) „thousands“ bzw. „sections“. Eine aus mindestens drei Ziffern[13] bestehende DDC-Ziffernfolge setzt sich folgendermassen zusammen: die 1. Ziffer repräsentiert die Hauptklasse („tens“), die 2. Ziffer die Division („hundreds“) und die 3. Ziffer die Sektion („thousands“). Nach der 3. Stelle folgt immer ein „Dewey-Punkt“ („.“), ggf. nachfolgende Ziffern werden für weitere inhaltliche Verfeinerungen („subdivisions“) benutzt. So erhalten Werke über den freien Willen die DDC-Notation (DDC number) „123.5“ und gehören demnach zur 4. DDC-Hierarchie-Ebene:

	Main Classes
100	Philosophy & psychology
120	Epistemology
123	Determinism and indeterminism
123.5	**Freedom**

Abb. 2.2 Bsp. für 4 DDC-Hierarchie-Ebenen (aus WebDewey-Datensatz der Klasse „Freedom“[14])

Hauptlieferanten für DDC-Notationen in GVK-Titeldatensätzen entstammen BNB (British National Bibliography)- und LoC-Titeldatensätzen, sind aber auch in anderen (z. B. DDB, SBB-PK Berlin, ZDB)-Titeldatensätzen enthalten. In naher Zukunft wird es von DDB - als ein Ergebnis des Projektes DDC Deutsch - auch DDC-Notationen für deutsche Veröffentlichungen geben, z. B. für Online-Dissertationen. Ab 2004 werden von DDB ausserdem DDC-basierte Sachgruppen (vorrangig in die ersten 100 DDC-Klassen der 2. Ebene) vergeben[15]. D.h., ein Anteil der GVK-Titeldatensätze sind schon DDC-klassifiziert und dieser Anteil wird durch die Neuzugänge erheblich zunehmen. Es ist daher naheliegend, eine mit z. B. DDC vorgenommene Klassifizierung durchgängige Systematisierung des GVK-Titelbestandes für fachgebietszentrierte Fragestellungen vorzunehmen. Für den verbleibenden nicht DDC-klassifizierten Löwenanteil des GVK-Titelbestandes allerdings stellt sich die Frage:

„Kann Kolibri diesen automatisch nach DDC klassifizieren?“[16]

Die Suche nach einer Antwort o.g. Teilaufgabe (3), d. h. Objektklassifikation, hat begonnen und ist Inhalt folgender Kapitel.

[13] DDC-Notationen für Klassen der ersten beiden Hierarchie-Ebenen werden mit Nullen („0“) auf 3 Stellen ver-
vollständigt, z. B. wird die Hauptklasse „Literatur“ nicht mit „8“ sondern mit der Notation „800“ gekennzeichnet und die Klasse der zweiten Ebene „Fotographie, Computerkunst“ lautet nicht „77“, sondern „770“

[14] http://connexion.oclc.org/WebZ/QUERY?sessionid=sp05sw03.dev.oclc.org-57770-dhqc4632-gdihzb&termsrch-sn%3A=123.5&next=/WebZ/html/corc/corcframe.html:entitycorcLink=/WebZ/html/corc/deweytrecordframe.html&bad=html/corc/badsearch.html&:tdbname=DeweyDB:entityDeweyNumber=123.5

[15] http://www.ddc-deutsch.de/, (Auskunft von Herrn Dr. Lars Svensson vom 23.1.04)

[16] http://www.fantasten.de/afrika3.htm („Warum der Kolibri der König der Tiere ist“, s. Anhang)

3. Test- und Entwicklungsumgebung

Einen DDC-klassifizierten Zugang zu Internetpublikationen bieten "La Trobe University" und die "National Library of Canada"[17] an. Auch OCLC ist in Forschungsprojekten zur automatischen Klassifikation engagiert:

"Several research projects are underway to explore automatic application of classification. OCLC's Scorpion project [13] and the University of Wolverhampton's automatic classifier [14] are two such attempts. Neither approach is able to fully exploit the capabilities of the underlying system by number-building. Research using Liu's thesis on decomposition [15] will not only aid subject analysis (see above), but also support automatic number-building." [18]

Das OCLC-Indexierungs- und Katalogisierungsprojekt „Scorpion"[19] ist bspw. für das VZG-Projekt „Colibri" untersuchenswert, da es umfangreich dokumentiert, gut fundiert[20] und der Scorpion-Quellcode frei verfügbar ist[21]. Als offenes Softwareprojekt konzipiert, kann mit einer CVS (Concurrent Versions System)[22]-Benutzerkennung an der Scorpion-Softwareentwicklung partizipiert werden. Projekte wie „CARMEN", „DESIRE", „GERHARD", „KASCADE", „MILOS", „OSIRIS", „Renardus" sind sicherlich auch betrachtenswert, doch der verbleibende Teil dieses Berichtes soll dem Stand des VZG-Projektes „Colibri" vorbehalten bleiben. Es folgt die Beschreibung der VZG-Colibri-Test- und Entwicklungsumgebung.

Der Prototyp „COLIBRI-DDC" wird auf einem Subnotebook / PC (1.2 GHz, 256 KB / 1 GB, 20 GB / 60 GB Plattenspeicher) unter Linux (Kernel 2.4.20-4GB) in der Sprache „awk" unter Zuhilfename üblicher Unix-(Shell)-Kommandos wie „bash", „cut", „cat", „head", „ksh", „split", „tail", „time" etc. realisiert. Der Name „awk" entstammt den Anfangsbuchstaben seiner Schöpfer: Alfred V. Aho, Peter J. Weinberger und Brian W. Kernighan (C-Sprachentwickler).

---

[17] http://library.bendigo.latrobe.edu.au/irs/webcat/ddcindex.htm,
http://www.nlc-bnc.ca/caninfo/ecaninfo.htm

[18] http://www.oclc.org/dewey/research/research_agenda.htm,
13. Shafer, Keith, Subramanian, Srividhya, and Fausey, Jon. "Measures for Evaluating Automatic Subject Assignment of Electronic Resources." 1999. Available at: http://orc.rsch.oclc.org:6109/measures.html
14. Jenkins, Charlotte, Jackson, Mike, Burden, Peter, and Wallis, Jon. "Automatic RDF Metadata Generation for Resource Discovery." Paper presented at the 8th International World Wide Web Conference, May 11-14, 1999. Available at: http://www8.org/w8-papers/2c-search-discover/automatic/automatic.html
15. Liu, Songqiao. "The Automatic Decomposition of DDC Synthesized Numbers." Ph.D. diss., University of California, Los Angeles, 1993. (Zusatz d. Autorin: Liu, Songqiao: „Decomposing DDC Synthesized Numbers". Paper presented at the 62nd IFLA General Conference, Beijing. ICBC, Vol. 26, No. 3, July/Sept. 1997, pp. 58-62, Online-Version: http://www.ifla.org.sg/IV/ifla62/62-sonl.htm)

[19] http://orc.rsch.oclc.org:6109/

[20] stützt sich u.a. auf Salton's Ergebnisse, Wegbereiter der experimentellen Forschung im Gebiet des Information Retrieval (http://www.asis.org/Features/Pioneers/salton.htm), WWW-Pionier Tim Berners-Lee und weitere

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

[22] http://www.cvshome.org/

Seit der 1. Version „oawk“ („old awk“) aus dem Jahr 1977 gibt es seit 1987 inzwischen „nawk“ („new awk“)[23]. In COLIBRI-DDC wird z. Zt. unter Linux die Variante „gawk“ (Implementierung der Programmiersprache „awk“ des GNU-Projektes, „POSIX 1003.2 Command Language And Utilities Standard“-konform, Version 3.1.1) verwendet. Die Programmiersprache „awk“ ist für Aufgaben, bei denen Textmuster-Suche und -Manipulationen im Vordergrund stehen und für eine effiziente Prototyp-Entwicklung gut geeignet. Falls eine Effizienzsteigerung der Laufzeit notwendig wird, kann COLIBRI-DDC in eine andere Sprache übersetzt werden; direkt oder ev. mit Übersetzungsprogrammen wie „awkcc“, „awk2c“ oder „a2p“ (Übersetzung von „awk“ nach „C“ oder „Perl“). Ein awk-Programm ist eine Anweisungsfolge der Form:

Muster {Aktion}
Muster {Aktion}
.
.
.

und optionalen Funktionsdefinitionen. „awk“ stellt u.a. Ein-/Ausgabe-, numerische, Zeit-, Zeichenketten- und Bitmanipulations-Funktionen zur Verfügung. Der aktuelle Implementierungs-Stand zu COLIBRI-DDC (in Bash- / Kornshell-Skripte eingebettete awk-Programme) befindet sich im Anhang.

Einfache mathematische Modellbildung (s. Anhang „ul-co-mod1“) und eine explorative Vorgehensweise wurden zunächst eingeschlagen, um möglichst schnell zu ersten Erkenntnissen / Ergebnissen zu kommen. Ausgangspunkt bilden GVK-Titeldatensätze, die mindestens eine DDC-Notation (ddc_no) enthalten. Diese werden mit dem DDC-Vorverarbeitungs-Programm „ul-ddc-pre“ („pre“ steht für „preprocessing“ oder „pre-paration“) in eine – für die daran anschliessende DDC-Klassifizierung (mit DDC-Programm „ul-ddc“) vorteilhafte – deskriptor-orientierte Form gebracht. Unter „Deskriptor“ wird hier jedes, den Kategorien entnommene, (inhalts-)beschreibende Element, verstanden. Die (Lang)-Repräsentation eines Deskriptors ist ein 8-Tupel folgenden Aufbaus:

KKKK-descr:= ddc_no|descr_val|tag|ind|ind_val|pos|pub_year|ppn

wobei: „KKKK-descr“: Deskriptorname mit „KKKK“: 4-stell. Kategoriennummer; ddc_no: DDC Notation (number); descr_val: Deskriptorwert (mehrere Wörter werden je nach Kategorie durch „_“ bzw. „#“ verbunden); tag: Kategorie; ind: Indikator; ind_val: Indikatorwert; pos: Position des Deskriptorwertes; pub_year: Veröffentlichungsjahr; ppn: Pica Produktionsnummer. Nicht vorhandener Deskriptorwert wird durch „_“ dargestellt. Großbuchstaben werden in Kleinbuchstaben (mit „tolower“) umgewandelt. Die Menge aller „KKKK-descr“ bilden die DDC-Basis („ddc_base“, s. Abb. 3.1). Tupelelemente werden durch das - auch von Pica verwendete - Trennzeichen „florin“ (Oktal „\237“) voneinander getrennt (hier als „Ÿ“ zu lesen). Beispiel für einen 044A-Deskriptor aus der DDC-Basis (ddc_base long):

123.5092Ÿ#p#locke_john#1632-1704#contributions_in_free_will_and_determinism
Ÿ044GŸ_Ÿ_Ÿ_Ÿ_Ÿ346746906

[23] Helmut Herold: Linux – Unix – Profitools. awk, sed, lex, yacc und make. 3. überarb. Aufl. Addison-Wesley, Bonn u.a., 1999

Derzeit besteht die DDC-Basis aus Deskriptoren der folgenden Kategorien aller GVK-Titeldatensätze, deren DDC-Notation mit 0, 1, 2, ... , 9 beginnen: „Hauptsachtitel“ (021A), „1.-3. Verfasser“ (028A, 028B), „sonstige beteiligte Personen“ (028C), „Verknüpfung zur größeren Einheit” (039B), „Library of Congress Subject Headings“ (044A), „British Library Subject Headings“ (044G), „Einzelschlagworte“ (044K), „Library of Congress Classification” (045A) und „Basisklassifikation“ (045Q). Der auf angegebene Weise selektierte DDC-Titeldatenbestand umfasst im GVK (Stand: April 2003) insgesamt ca. 2.8 Mio. DDC-klassifizierte Datensätze, wobei sich diese wie folgt auf die hier mit „ddc 0*“ bis „ddc 9*“ bezeichneten Notationen verteilen:

geordnet nach Klassennummer		geordnet nach Anzahl Titelsätzen	
ddc 0*	125.742	ddc 4* 59.759	2.1 %
ddc 1*	77.227	ddc 1* 77.227	2.8 %
ddc 2*	148.934	ddc 0* 125.742	4.5 %
ddc 3*	690.241	ddc 2* 148.933	5.3 %
ddc 4*	59.759	ddc 5* 175.551	6.2 %
ddc 5*	175.533	ddc 7* 259.885	9.2 %
ddc 6*	411.131	ddc 9* 363.396	12.9 %
ddc 7*	259.885	ddc 6* 411.131	14.6 %
ddc 8*	503.990	ddc 8* 503.990	17.9 %
ddc 9*	363.396	ddc 3* 690.241	24.5 %

Tabelle 3.1 Prozentuale Verteilung der GVK-Titeldatensätze in den DDC-Wissensgebieten

Die zehn Hauptklassen (Erste Übersicht) sind:

000	Informatik, Informationswissenschaft, allgemeine Werke
100	Philosophie und Psychologie
200	Religion
300	Sozialwissenschaften
400	Sprache
500	Naturwissenschaften und Mathematik
600	Technik, Medizin, angewandte Wissenschaften
700	Künste und Unterhaltung
800	Literatur
900	Geschichte und Geografie

Tabelle 3.2 Notation (Klassennummern) und Bezeichnungen der DDC-Hauptklassen

Die in Tabelle 3.1 angegebenen GVK-Titeldatensatzmengen ddc 0*, ... , ddc 9* wurden mithilfe der WinIBW-Kommandos „finde“ („f ddc 0*“, … , „f ddc 9*“) und „download“ („dow s1 p“, … , „dow s10 p“)[24] erzeugt und in zehn Dateien abgespeichert. Die ungefähren Mengenverhältnisse[25] ddc- und nicht-ddc-klassifizierter Titeldatensätze im GVK-PLUS sind folgender Abbildung zu entnehmen:

[24] „s1“, ... , „s10“ steht für Selektionsmenge 1, ... , 10 und „p“ für internes Datenformat (Pica+ Präsentation)

[25] Das Universum müsste wesentlich größer dargestellt werden.

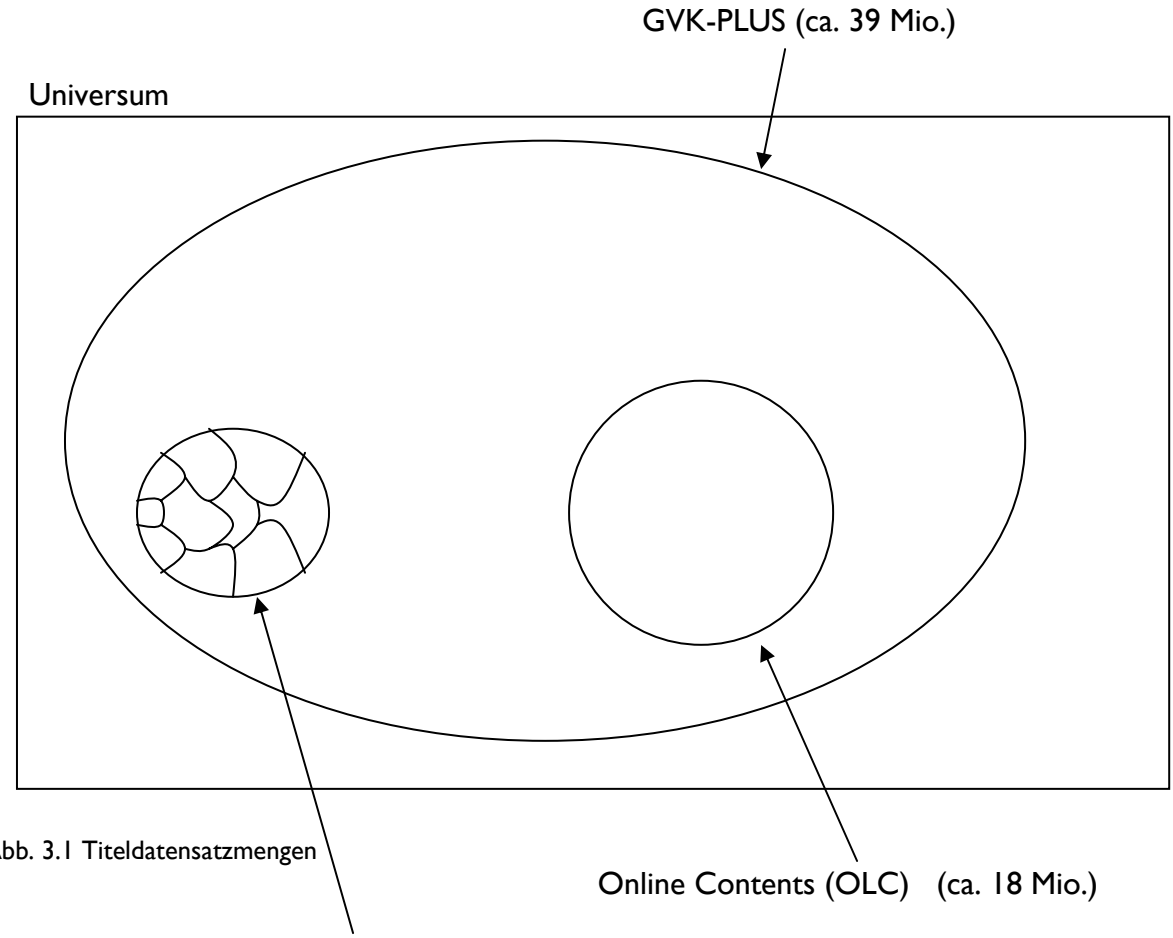


Abb. 3.1 Titeldatensatzmengen

Da die im GVK enthaltenen DDC-Notationen von Fachreferenten vergeben wurden, ist die DDC-Basis eine empirische Basis. Diese intellektuell erstellte Basis stellt Information für die automatische Klassifikation bereit. Im Laufe des Projektes soll untersucht werden, ob diese empirische Basis (z. B. Größe, Qualität) für eine inhaltlich stimmige automatische Klassifikation ausreicht oder durch weitere Information angereichert werden muß: Berücksichtigung weiterer Kategorien wie „Medical Subject Headings“ (044C), „PRECIS“ (044E), „DDB-Schlagwörter“ (044F), „Angaben zum Inhalt“ (046M), „Inhaltliche Zusammenfassung“ (047I), weitere ddc-klassifizierte (Titel)-Bestände, Beziehungen zwischen Begriffen oder Verwendung / Erstellung von Konkordanzen, automatische Thesauruskonstruktion, etc.

Zusätzlich zur 8-Tupel-Darstellung der Deskriptoren (ddc_base long), in der alle wesentliche Information (z. B. auch Wortposition von Titelwörtern) enthalten ist, wurde für erste Tests aus Effizienzgründen eine 2. Darstellungsform, eine Kurzform der DDC-Basis (ddc_base short), hergestellt. Die Repräsentation eines Deskriptors wird zum 2-Tupel:

KKKK-descr:= ddc_no|descr_val

Der oben angegebene Deskriptor der „ddc_base short“ ergibt sich zu:

123.5092Ÿ#p#locke_john#1632-1704#contributions_in_free_will_and_determinism

Die „ppn" ist in der Kurzform auch eliminiert, da es für die Bildung der DDC-Basis „ddc_base short" (im Folgenden kurz: „ddc_base") unerheblich ist, welche „ppn" mit einem Deskriptorwert einer Kategorie zu einer DDC-Basis-Klasse beiträgt. Folgende Kommandofolge erzeugt aus der 8-Tupel-Darstellung die 2-Tupel-Darstellung der DDC-Basis:

```
#!/bin/ksh
tr "\237" "\040" < $1 | cut -f 1,2 -d" " | sort | uniq | tr "\040" "\237" > $1_uniq
```

Abb. 3.2: Korn-Shell-Skript „ul-shorten-ddc_base"

Nach Kürzung auf ein 2-Tupel wird nach DDC-Klassennummern aufsteigend sortiert („sort") und mehrfaches Vorkommen eines Deskriptorwertes auf ein einmaliges Vorkommen (da die DDC-Basis als Menge realisiert wird) reduziert („uniq"). Danach werden mit dem Korn-Shell-Skript „ul-revise-ddc_base" Zeilen eliminiert, die - aus unterschiedlichen Gründen - mit inkorrekten DDC-Notationen beginnen:

```
#!/bin/ksh
cat $1 | grep -v "^[^0-9]" | grep -v "^.\{1,2\}\>" > $1_rev
```

Abb. 3.3: Korn-Shell-Skript „ul-revise-ddc_base"

Da manche Titeldatensätze mit mehr als einer DDC-Notation versehen sind, kann es vorkommen, daß DDC-Notationen in mehreren der zehn Einzeldateien „file_ddc0_base", ... „file_ddc9_base" der DDC-Basis auftreten. Aus Abbildungen Abb. 3.4 und Abb. 3.5 ist z. B. zu erkennen, daß die DDC-Notation „999" („Extraterrestrial worlds") in unterschiedlichen Teilen der DDC-Basis („file_ddc0_base", „file_ddc5_base", „file_ddc9_base") vorkommt:

```
000Ÿ#a#andes_region#x#antiquities
000Ÿ#a#architecture
000Ÿ#a#cycling
000Ÿ#a#cyclists#z#england#x#biography
000Ÿ#a#dartmoor_(england)#x#guidebooks
000Ÿ#a#incas#x#antiquities
000Ÿ#a#moreau_de_maupertuis#h#pierre-louis
000Ÿ#a#pottery#z#andes
000Ÿ#a#pottery_ancient#z#andes
000Ÿ#a#science
...
998/.9Ÿg860
998/.9Ÿphilip_w.#quigg
998/.9Ÿthe
999Ÿ#g#antarctica#politics_and_government
999Ÿaffairs
999Ÿantarctica
999Ÿe744
999Ÿin
999Ÿlaurence_mckinley#gould
999Ÿworld
```

Abb. 3.4: Auszug (erste 10 und letzte 10 Zeilen) aus DDC-Basis "file_ddc0_base_first_10_last_10"

Um diesem Phänomen zu begegnen, wird die DDC-Basis nochmals überarbeitet. Das Korn-Shell-Skript „ul-make-ddc_base" (Abb. 3.6) stellt aus den Einzeldateien die DDC-Basis in einer einzigen Datei „file_ddc_base" zusammen („cat"), sortiert und teilt diese wieder in 10 Einzeldateien „file_ddc0_base" (1. Teildatei der DDC-Basis), „file_ddc1_base" (2. Teildatei

der DDC-Basis) ..., „file_ddc9_base“ (10. Teildatei der DDC-Basis) auf. Wie aus Abb. 3.5 zu sehen, sind noch irrelevante Deskriptorwerte wie „in”, „from”, „to” und „the” in der DDC-Basis enthalten, die mit dem Korn-Shell-Skript „ul-make-elim” (Abb. 3.7) beseitigt werden.

```
file_ddc0_base:999Ÿ#g#antarctica#politics_and_government
file_ddc0_base:999Ÿaffairs
file_ddc0_base:999Ÿantarctica
file_ddc0_base:999Ÿe744
file_ddc0_base:999Ÿin
file_ddc0_base:999Ÿlaurence_mckinley#gould
file_ddc0_base:999Ÿworld
file_ddc5_base:999Ÿ#a#science#x#history
file_ddc5_base:999Ÿ#a#science#x#philosophy
file_ddc5_base:999Ÿ#b#extraterrestrial_life
file_ddc5_base:999Ÿ02.01_geschichte_der_wissenschaft_und_kultur
file_ddc5_base:999Ÿ30.01_geschichte_der_naturwissenschaften
file_ddc5_base:999Ÿau¾erirdisches_leben
file_ddc5_base:999Ÿcopernican
file_ddc5_base:999Ÿfiction
file_ddc5_base:999Ÿfrom
file_ddc5_base:999Ÿfrontier
file_ddc5_base:999Ÿgeistesgeschichte
file_ddc5_base:999Ÿhelen#atkins
file_ddc5_base:999Ÿimagining
file_ddc5_base:999Ÿkarl_s.#guthke
file_ddc5_base:999Ÿlast
file_ddc5_base:999Ÿliteratur
file_ddc5_base:999Ÿmodern
file_ddc5_base:999Ÿnaturwissenschaften
file_ddc5_base:999Ÿother
file_ddc5_base:999Ÿq125
file_ddc5_base:999Ÿrevolution
file_ddc5_base:999Ÿscience
file_ddc5_base:999Ÿthe
file_ddc5_base:999Ÿto
file_ddc5_base:999Ÿworlds
file_ddc9_base:999Ÿ#a#antarctica
file_ddc9_base:999Ÿ#a#antarctica#x#history
file_ddc9_base:999Ÿ#a#astronomy#x#popular_works
file_ddc9_base:999Ÿ#a#discovery_(ship)
file_ddc9_base:999Ÿ#a#exobiology
file_ddc9_base:999Ÿ#a#explorers#z#antarctica#x#diaries
file_ddc9_base:999Ÿ#a#interstellar_communication
file_ddc9_base:999Ÿ#a#lashly_william#x#diaries
file_ddc9_base:999Ÿ#a#life_on_other_planets
...
```

Abb. 3.5: Vorkommen der DDC-Notation „999“ in den Einzeldateien „file_ddc0_base“, „file_ddc5_base“ und „file_ddc9_base“ der DDC-Basis

```
#!/bin/ksh
cat file_ddc*_base_uniq > file_ddc_base;
sort –u file_ddc_base > file_ddc_base-sorted
split -a1 --bytes=52m file_ddc_base-sorted file_ddc_base;
mv file_ddc_basea file_ddc0_base;
...
mv file_ddc_basej file_ddc9_base;
```

Abb. 3.6: Korn-Shell-Skript „ul-make-ddc_base”

```
#!/usr/bin/ksh
# ul, 6.12.03
for i in file_ddc[0-9]_base; do
  print $i
  ~/colibri/ul-awk/ul-elim $i | sort –u > ${i}_elim_uniq
done
```

Abb. 3.7: Korn-Shell-Skript „ul-make-elim"

Derzeit werden folgende Deskriptorwerte eliminiert:

```
...
ELIM_1_SPEC_CHAR= "\\(|\\)|\\[|\\]|\\^|'|´|`|?|!|\"|\\.|,|:|;|<|>|-|+|&"
ELIM_1CHAR      = "[a-z0-9]"
ELIM_2CHAR      =  "ad|al|an|as|at|be|by|da|di|de|do|du|el|en|et|he|if|ii|il|im|in|is|it|la|le|li|lo|mo|na|nh|od|of|
                    on|or|ou|to|up|vo|we|xi|zu"
ELIM_3CHAR      = "all|als|and|any|are|aus|bis|can|das|del|dem|den|der|des|die|dun|ein|for|his|how|iia|iic|iid|
                   iie|iii|iis|its|las|les|our|que|sin|the|und|una|une|vii|vom|von|you|zum|zur|was|xii"
ELIM_4CHAR      = "from|hrsg|iiav|iics|iigs|iiii||iisg|iisi|jiia|less|mkii|more|nach|oder|with|that|what|
                    your|viii|xiii|xvii"
ELIM_5CHAR      = "della|delle|iiiii|ixiii|viiie|xviii|xxiii"
ELIM_6CHAR      = "anyone|viiiix|xviiic|xviiie"
ELIM_7CHAR      = "vixviii|xvxviii"
ELIM_8CHAR      = "zwischen"
...
```

Abb. 3.8: Auszug aus awk-Programm „ul-elim" (gesamtes Programm im Anhang)

Nach allen Bearbeitungsschritten umfasst die gesamte DDC-Basis „file_ddc_base_elim_uniq" 505 Megabyte (MB), in den 10 Teilen für die Hauptspeichergröße angepaßte Größen zwischen 46 und 52 MB (s. Kapitel 4):

```
short/elim> ls -lh
insgesamt 1011M
-rw-r-----   1 ul      users        51M 2003-12-12 18:11 file_ddc0_base_elim_uniq
-rw-r-----   1 ul      users        52M 2003-12-12 18:13 file_ddc1_base_elim_uniq
-rw-r-----   1 ul      users        51M 2003-12-12 18:13 file_ddc2_base_elim_uniq
-rw-r-----   1 ul      users        52M 2003-12-12 18:14 file_ddc3_base_elim_uniq
-rw-r-----   1 ul      users        51M 2003-12-12 18:15 file_ddc4_base_elim_uniq
-rw-r-----   1 ul      users        52M 2003-12-12 18:16 file_ddc5_base_elim_uniq
-rw-r-----   1 ul      users        51M 2003-12-12 18:20 file_ddc6_base_elim_uniq
-rw-r-----   1 ul      users        52M 2003-12-12 18:20 file_ddc7_base_elim_uniq
-rw-r-----   1 ul      users        51M 2003-12-12 18:21 file_ddc8_base_elim_uniq
-rw-r-----   1 ul      users        46M 2003-12-12 18:21 file_ddc9_base_elim_uniq
-rw-r-----   1 ul      users       505M 2003-12-12 17:10 file_ddc_base_elim_uniq
```

Abb. 3.8: ddc_base „file_ddc_base_elim_uniq" und ihre Teile

Nach Präparierung der DDC-Basis wird die der Titeldatensätze vorgenommen: Während DDC-klassifizierte Titeldatensätze mit „ul-ddc-pre infile yes" (a) hergestellt werden, werden zu DDC-klassifzierende Titeldatensätze mit „ul-ddc-pre infile no" (b) präpariert, wobei „infile" die Eingabedatei der (klassifzierten / zu klassifizierenden) n[26] Original-Titeldatensätze ist und „yes"/„no" die Frage, ob es sich um eine Datei der DDC-Basis handelt, beantwortet. Abb. 3.9 zeigt einen mit (a) erzeugten Teil der DDC-Basis (Titeldatensatz PPN 356730425), in Abb. 3.10 und Abb. 3.11 befinden sich mit (b) erzeugte Titeldatensätze. Nicht inhaltstragende Wörter werden wieder mit „ul-elim" vor Ausgabe eliminiert und Autoren in

[26] n = (1, ..., max.) Titeldatensatznummer

der Form: „Vorname#Nachname“ ausgegeben. Aus DDC_NO-Wert „XXX“ (Abb. 3.10) ist abzulesen, daß der Titel nicht ddc-klassifiziert ist. Variante (b) kann auch für DDC-klassifzierte Titeldatensätze benutzt werden[27]. In diesem Fall steht statt „XXX“ die vom Fachpersonal dem Titel zugeordnete DDC-Notation „006.696” („Digital video“).

```
006.696Ÿdenise#seguinŸ028AŸd.aŸ_Ÿ_Ÿ2004Ÿ356730425
006.696ŸmacromediaŸ021AŸaŸ_Ÿ1Ÿ_Ÿ356730425
006.696ŸflashŸ021AŸaŸ_Ÿ2Ÿ_Ÿ356730425
006.696ŸmxŸ021AŸaŸ_Ÿ3Ÿ_Ÿ356730425
006.696Ÿcomputer_animationŸ044AŸ_Ÿ_Ÿ_Ÿ_Ÿ356730425
006.696Ÿweb_sites#designŸ044AŸ_Ÿ_Ÿ_Ÿ_Ÿ356730425
006.696Ÿflash_(computer_file)Ÿ044AŸ_Ÿ_Ÿ_Ÿ_Ÿ356730425
006.696Ÿtr897.7Ÿ045AŸaŸ_Ÿ_Ÿ_Ÿ356730425
```

Abb. 3.9 Mit „ul-ddc-pre infile yes“ hergestellter Teil der DDC-Basis „ddc_base_long_ppn_356730425“

```
implementation
object-oriented
multimedia
knowledge
base
management
system
model
language
h.#ishikawa
f.#suzuki
f.#kozakura
a.#makinouchi
m.#miyagishima
y.#izumida
m.#aoshima
y.#yamane
association for computing machinery: acm transactions on database systems
DDC_NO= XXX
PPN= 50328596x
```

Abb. 3.10 Präparierter, zu DDC-klassifizierender OLC-Titeldatensatz (file_to_ddc_olc_50328596x)

```
macromedia
flash
mx
denise#seguin
flash_(computer_file)
computer_animation
web_sites#design
tr897.7
DDC_NO= 006.696
PPN= 356730425
```

Abb. 3.11 Für einen Qualitätstest der DDC-Klassifikation präparierter, klassifizierter Titeldatensatz

Mit „ul-ddc file_ddc_base file_to_ddc“ (Programm „ul-ddc“ s. Anhang) kann nun klassifiziert werden. Die durch „file_ddc_base“ angegebene Datei der DDC-Basis wird in ein Feld (array) vollständig[28] eingelesen, jeder Deskriptorwert („$1“) des zu klassifizierenden Titeldaten-

[27] z. B. für Testzwecke (s. Kap. 4, Bsp. 5, S. 21)

[28] Die max. Größe des Teils der einzulesenden DDC-Basis hängt von der Hauptspeichergröße ab.

satzes (oder mehrerer zu klassifizierender Titeldatensätze) wird mit jedem Deskriptorwert („descr_val“) der DDC-Basis verglichen. Wenn die Zeichenkette „$1“ in der Zeichenkette „descr_val“ enthalten ist (Variante 1), wird der Zähler der DDC-Klasse, zu der der „descr_val“ gehört, um eins erhöht[29] und am Ende wird die DDC-Klasse mit der größten Häufigkeit ausgegeben. Anstelle von (Variante 1) kann auch auf exakte Übereinstimmung der Deskriptorwerte (Variante 2) geprüft oder z. B. mit einer kontextabhängigen Variante-1-2-Kombination experimentiert werden. Das beschriebene und z. Zt. implementierte Verfahren realisiert das in der Literatur bekannte, einfachste Ähnlichkeitsmaß, das Vektorkorrelationsmaß[30] (auch Skalarprodukt genannt), das benutzt werden kann, um die Ähnlichkeit α zwischen Titeldatensätzen, denen Eigenschaftsvektoren der Länge l(ength) zugeordnet sind, zu bestimmen:

$$\alpha_{bt} = \sum_{i=1}^{l} b_i t_i \qquad (\alpha_{\text{Skalarprodukt}})$$

wobei:
$b \in B$ (B: Menge der klassifizierten Titeldatensätze $\cong$ DDC-Basis)
$t \in T$ (T: Menge der zu klassifizierenden Titeldatensätze)

Sei
b = (1,0,1,1,0,0,1,0,1,1,0)
t = (0,1,1,0,0,0,1,1,1,1,1)

mit
1: Deskriptorwert vorhanden
0: Deskriptorwert nicht vorhanden

Dann ergibt $\alpha_{\text{Skalarprodukt}} = 4$.

Es gibt viele weitere Korrelationsmaße, mit denen experimentiert werden kann, wobei jedes seine Stärken und Schwächen hat. In die Maße gehen unterschiedliche Eigenschaften ein, z. B. Eigenschaften, die zwei Titeldatensätze gemeinsam haben (wie bei „$\alpha_{\text{Skalarprodukt}}$“) oder nicht gemeinsam haben. Dies wird durch binäre Gewichtungen dargestellt. Verfeinerungen berücksichtigen Normierungen, Verhältnis der Längen, Asymmetrien, usw. Bekannte Maße sind z. B. „Tanimotomaß“, „Cosinuskoeffizient“, „Überlappungsmaß“, „Maß von Maron und Kuhns“, „Maß von Bennett und Spiegel“. Um herauszufinden, welche Maße sich eignen könnten, muß die DDC-Basis und der Titeldatenbestand GVK-PLUS analysiert werden. Ein erster Analyseschritt ist die Ermittlung der Häufigkeiten von Deskriptorwerten (s. Anhang „ul-freq“). Mit Tests, bei denen unterschiedliche Klassifikationsverfahren eingesetzt werden, soll versucht werden, die eingangs gestellte Frage Q1 zu beantworten.

[29] ddc_freq_array[ddc_no]++

[30] Gerard Salton: Automatic Information Organization and Retrieval. Chapter 7-2: Association Coefficients for Term and Document Vectors. McGraw-Hill, Inc., New York u.a., 1968, pp. 236

4. Ergebnisse

Die Praxis soll das Ergebnis des Nachdenkens sein, nicht umgekehrt.
(Hermann Hesse)

Aus Gründen der Machbarkeit (max. verfügbare Hauptspeichergröße) wird die gesamte DDC-Basis (2.8 Mio. Titeldatensätze) - wie in Kapitel 3 beschrieben - in 10 einzelne Dateien („file_ddc0_base“, ... , „file_ddc9_base“) aufgeteilt, deren Klassifizierungsergebnisse am Ende zusammengeführt werden. Dies wird mit folgender Kommandofolge „ul-match“ realisiert:

```
#!/usr/bin/ksh
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc0_base_elim_uniq $1
echo "file_ddc0_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc1_base_elim_uniq $1
echo "file_ddc1_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc2_base_elim_uniq $1
echo "file_ddc2_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc3_base_elim_uniq $1
echo "file_ddc3_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc4_base_elim_uniq $1
echo "file_ddc4_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc5_base_elim_uniq $1
echo "file_ddc5_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc6_base_elim_uniq $1
echo "file_ddc6_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc7_base_elim_uniq $1
echo "file_ddc7_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc8_base_elim_uniq $1
echo "file_ddc8_base_elim_uniq done"
/home/ul/colibri/ul-awk/ul-ddc /home/ul/colibri/ddc_base/short/elim/file_ddc9_base_elim_uniq $1
echo "file_ddc9_base_elim_uniq done"
sort -k1,1 -k4,2n top* > toto  # top total ddc candidat
echo "DDC-Notation: ..."
tail -1 toto
```

Abb. 4.1 „ul-match“: DDC-Klassifizierung der in Datei „$1“ enthaltenen Titeldatensätze

Mit den in Kapiteln 3 und 4 beschriebenen und im Anhang enthaltenen Programmen und Korn-Shell-Skript-Kommandofolgen steht nun eine (leicht erweiterbare) COLIBRI-DDC-Test- und Entwicklungsumgebung zur Verfügung, mit der begonnen werden kann, zu experimentieren.

Es sei folgender Titeldatensatz „infile_to_ddc_hummingbird“ gegeben (Bsp. 1):

```
hummingbird
DDC_NO= XXX
PPN= 123456789
```

Abb. 4.2 Fiktiver Titeldatensatz „infile_to_ddc_hummingbird“

Der Titeldatensatz enthält einen einzigen Deskriptorwert, der z. B. der Titelkategorie („021A“) entstammen kann. Der Aufruf „ul-match infile_to_ddc_hummingbird“ erzeugt in ca. 4 Minuten die DDC-Notation „598.8/99“. Eine in WebDewey gestellte Anfrage „Search 598.8 in Dewey Numbers“ liefert folgende Systemantwort:

	Main Classes
500	Science
579-590	Natural history of specific kinds of organisms
580-590	Plants and animals
590	Animals (Zoology)
592-599	Specific taxonomic groups of animals
598	*Aves (Birds)
598.5-598.9	Land birds
598.8	***Passeriformes (Perching birds)**

Abb. 4.3 Auszug aus WebDewey-Datensatz „Passeriformes (Perching birds)“

und die WebDewey-Anfrage „Search hummingbird* in Relative Index“ ergibt:

3 records found searching for ri: (hummingbird*) in WebDewey:

1.	B	333.958764	Hummingbirds--resource economics
2.		598.764	*Trochili (Hummingbirds)
3.	B	639.978764	Hummingbirds--conservation technology

Abb. 4.4 Systemantwort zur WebDewey-Anfrage: „ri: (hummingbird*)“

Ein passendes Ergebnis, das von der automatischen Klassifikation (Variante 1, S. 15) geliefert wird, nämlich die DDC-Notation „598“[31] („Aves (Birds)“).

Beim nächsten Beispiel geht es um „Tauchen nach Fischen in Cadaqués” (Bsp. 2):

```
diving
fish
cadaques
DDC_NO= XXX
PPN= 9876543210
```

Abb. 4.5 Fiktiver Titeldatensatz „infile_to_ddc_diving“

Der Aufruf “ul-match ... “ erzeugt:

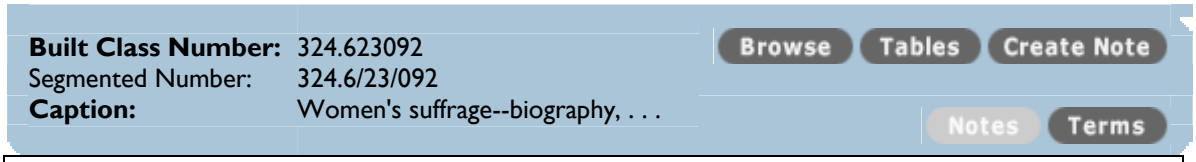
```
new/divi> time ~/colibri/ul-awk/ul-match infile_to_ddc_diving
 *** record done! ***file_ddc0_base_elim_uniq done
 *** record done! ***file_ddc1_base_elim_uniq done
 *** record done! ***file_ddc2_base_elim_uniq done
 *** record done! ***file_ddc3_base_elim_uniq done
 *** record done! ***file_ddc4_base_elim_uniq done
 *** record done! ***file_ddc5_base_elim_uniq done
 *** record done! ***file_ddc6_base_elim_uniq done
 *** record done! ***file_ddc7_base_elim_uniq done
 *** record done! ***file_ddc8_base_elim_uniq done
 *** record done! ***file_ddc9_base_elim_uniq done
DDC-Notation: ...
9876543210 XXX 799.1/755 192
330.740u 10.270s 5:43.71 99.2%  0+0k 0+0io 5764pf+0w
```

Abb. 4.6 Klassifizierungsausgabe von „ul-match ... infile_to_ddc_diving“

[31] nach Reduktion auf die ersten 3 Stellen (vgl. Q1-COLIBRI, S. 3), d.h. auf die ersten 1000 Klassen

Das COLIBRI-DDC-System[32] ermittelt für den Klassifizierungsprozess in ca. 6 Min. CPU-Zeit (bei 99.2% zugeteilter CPU)[33] die größte Ähnlichkeit[34] (192) für die DDC-Klasse „799.1/755“. In beiden Beispielen (Bsp. 1, Bsp. 2) ist in der DDC-Notation des Topkandidaten ein Schrägstrich „/“ enthalten, dessen Funktion nachfolgend erklärt wird.

Exkurs: Segmentierung
Mit einem Segmentierungszeichen „/“ („slash mark“), „ ’ “ („prime mark“) oder ggf. anderen Sonderzeichen wird eine Segmentierung gekennzeichnet[35]. Diese wird von zentralen Katalogdiensten wie der „LoC“ oder der „NLoC“[36] eingebracht, um die Position der Kürzungsmöglichkeit/en einer DDC-Notation zu kennzeichnen. Segmentierungszeichen können das Ende einer gekürzten DDC-Notation[37] oder den Beginn einer aus der DDC-Tabelle 1 („Standard Subdivision“) stammenden Notation kennzeichnen, z. B. [38]



	Main Classes
300	Social sciences
320	Political science
324	The political process
324.6	Election systems and procedures; suffrage
324.62	Suffrage
324.623	Women's suffrage
324.623092	**Women's suffrage--biography, . . .**

Abb. 4.7 Auszug aus WebDewey-Datensatz „Women’s suffrage--biography, ...“

Die Interpretation der ermittelten Klassifikationsergebnisse ergibt für „598.8/99“ (Bsp.1) „Passeriformes (Perching birds)-- Extraterrestrial worlds“, und für „799.1/755“ (Bsp. 2) die Klassenbenennung „Game fishes (Salmonids)--sports fishing, ...“. Die aus Bsp. 2 ermittelten anderen DDC-Notationen mit ihren Ähnlichkeiten sind:

```
9876543210 XXX 941.1 7
9876543210 XXX 248.4 12
9876543210 XXX 330 12
9876543210 XXX 428.6 20
9876543210 XXX 347.3037692 30
9876543210 XXX 823.914 81
9876543210 XXX 597/.58 142
9876543210 XXX 333.95/6 152
9876543210 XXX 639.3 162
9876543210 XXX 799.1/755 192
```

Abb. 4.8 Per COLIBRI-DDC-System ermittelte Klassen(häufigkeiten) für „infile_to_ddc_diving“

[32] wiederum mit Variante (1) von S. 15
[33] mit Linux-Kommando „time“ erzeugt
[34] Skalarprodukt aus Kapitel 3
[35] OCLC Connexion Help Center - DDC Appendix: Segmentation
[36] National Library of Canada
[37] „abridged number” (aus einer „abridged edition of the DDC”)
[38] Connexion Help Center - DDC Appendix: Segmentation

Auch die Klassen mit geringeren Ähnlichkeiten sind – je nach Kontext - inhaltlich passend:

639 Hunting, fishing, conservation, related technologies
639.3 Culture of cold-blooded vertebrates Of fishes
333 Economics of land and energy
333.95/6 Fishes
597 Cold-blooded vertebrates Pisces (Fishes)
597.5 Protacanthopterygii Salmoniformes (597/.58 ?ul?)
823 English fiction
823.914 English fiction--1945-1999, ...

Dies gibt Anlass zur folgenden Untersuchung (Bsp. 3):

```
diving
fish
cadaques
fiction
DDC_NO= XXX
PPN= 9999999999
```

Abb. 4.9 Fiktiver Titeldatensatz „infile_to_ddc_diving_fiction“

mit folgendem Ergebnis:

```
9999999999 XXX 944.034 10
9999999999 XXX 330 13
9999999999 XXX 016.823/0876 25
9999999999 XXX 347.3037692 30
9999999999 XXX 597/.58 142
9999999999 XXX 333.95/6 152
9999999999 XXX 741.5973 197
9999999999 XXX 428.6 1443
9999999999 XXX 823.914 12766
```

Abb. 4.10 Ermittelte DDC-Notationen (mit Ähnlichkeitswerten) für „infile_to_ddc_diving_fiction“

Das per Rechner erzielte Klassifikationsergebnis ist tatsächlich „823“ („English fiction“). Die zur Notation „823.914 ermittelte Ähnlichkeit von „12.766“ ist sehr groß. Ein Blick in die erstellten Worthäufigkeitstabellen „ddc_base_0_freq_sort“,...„ddc_base_9_freq_sort“ (Abb. 4.11)[39] zeigt, dass dieses Ergebnis auf die Häufigkeit[40] des Vorkommens des Wortes „fiction“ (27.101) in der Klasse „823.914“ zurückzuführen ist. Die Wörter „diving“ (10), „fish (6)“ und „cadaques[41](0)“ tragen nur sehr wenig bzw. das Wort „cadaques“ gar nicht dazu bei.

[39] erstellt mit „ul-freq“ (Ermittlung der Worthäufigkeiten der DDC-Basis „ddc_base_0“, ... „ddc_base9“)

[40] Im Text Größe dem Wort folgend in Klammern angegeben

[41] Das Wort „cadaques“ (mit Schreibvariante) kommt in der gesamten DDC-Basis nur zweimal in der Teilmenge ddc 7* und zweimal in der ddc 6*-Teilmenge vor: „cadaques“ einmal mit der zugeordneten DDC-Notation „709/.2/4“ („Historical, geographic, persons treatment of fine and decorative arts, Persons, Essays”), „cadaquâes“ zweimal mit DDC-Notation „709/.04“(„Historical, geographic, persons treatment of fine and decorative arts, Special topics”) und zweimal in „641.5946/7“ (“Cooking characteristic of specific geographic environments, ethnic cooking, Spain cooking, Eastern Spain and Andorra”)

070/.92 biography 206		523.01 congresses 148
025.04 computer 212		509 history 164
011/.31 catalogs 220		501 philosophy 169
005.8 computer 230		539.7 congresses 184
005.265 computer 249		501 science 194
005.13/3 computer 268		599/.0188 congresses 220
004.6 computer 280		500 congresses 261
005.1 computer 302		530.4/1 congresses 288
005.133 computer 342		510 mathematics 291
005.369 computer 348		500 science 305
174/.2 moral 217		629.28/722 repair 367
174/.2 ethical 220		629.28/722 maintenance 375
193 wilhelm 237		629.28722 automobile 391
194 philosophy 243		658 management 393
192 philosophy 249		629.28722 manuals 408
174/.2 aspects 250		629.2222 automobile 411
193 century 265		629.28/722 automobile 416
193 friedrich 288		610.73 nursing 423
193 history 290		629.28722 repair 424
193 philosophy 352		629.28722 maintenance 426
284.4 life 198		709 art 493
282.092 church 202		791.43/72 picture 493
282/.092 history 234		759.2 exhibitions 522
282/.092 catholic 275		709/.2 exhibitions 537
282.092 biography 282		791.4372 picture 538
282/.092 biography 319		791.43/72 motion 545
248.4 aspects 345		791.4372 motion 600
248.4 christianity 368		759 italy 627
282/.092 church 368		759.13 art 701
248.4 religious 402		759.13 exhibitions 780
378.73 united 511		813.6 fiction 3366
362.1 health 513		823.914 juvenile 3372
338.9 economic 593		823.92 fiction 3409
340 law 611		823.914 england 3541
362.1/0973 united 614		823.914 juvenile 3659
362.1/0973 states 635		823 fiction 3707
338.973 united 681		813/.54 fiction 9928
338.973 states 689		823.914 fiction 12679 !ul!
327.73 states 701		823.914 fiction 14372 !muss summiert werden!
327.73 united 774		813.54 fiction 19591
428 language 337		932 egypt 715
425 english 348		929/.2/0973 william 776
428.6 pictorial 365		929/.2/0973 john 995
428.6 works 370		973 states 1130
428 english 376		936.4 france 1192
428.6 readers 380		973 states 1198
428.6 literature 410		936.3 germany 1280
415 grammar 440		936.6 spain 1689
428.6 juvenile 1344		929.20973 family 1830
428.6 fiction 1422		929/.2/0973 family 12069

Abb. 4.11 Die10 häufigsten Wörter in „ddc_base_0_freq_sort“, ... „ddc_base_9_freq_sort“
(in der Reihenfolge: DDC-Notation Wort Vorkommenshäufigkeit)

Derartige und weitere (Häufigkeits-)Information ist für die Wahl des geeigneten Ähnlichkeitsmaßes und Klassifizierungsmodells aus den verschiedenen Gebieten wie Information Retrieval und Künstliche Intelligenz (Data /Text Mining) relevant.

Ein 1. (realer) OLC-Titeldatensatz (Bsp. 4)[42] soll nun mit dem COLIBRI-DDC-System klassifiziert werden. In ca. 16 Min. ermittelt das System mit Variante 1 als Topkandidaten „658“[43] („General management“) und mit einer geringeren Ähnlichkeit[44]die DDC-Notation „415“ (Grammar of standard forms of languages Syntax of standard forms of languages). Mit Variante 2, d.h. “exact match”, erzeugt das System in ca. 24 Min. die DDC-Notation „006.3“ („Artificial intelligence“) bzw. 371.9“ („Special education“), beide mit einem Ähnlichkeitswert „8“ und die DDC-Notation „658.8“ („General management“) mit dem Ähnlichkeitswert „7“. DDC-Experten mögen anhand des OLC-Titeldatensatzes selbst beurteilen, welche DDC-Notation vergeben werden müßte („006.3“ ist aus Sicht der Autorin eine geeignete Notation, ev. nicht die beste):

SET: S22 [157] TTL: 148 PPN: 50328596x SEITE1 .

001@ ŸaA
001A Ÿ02016:20-07-95
001B Ÿ02016:30-11-01Ÿt17:26:06.996
001D Ÿ09999:99-99-99
001X Ÿ02999
002@ Ÿ0Aox
011@ Ÿa1993
021A ŸaThe @Model, Language, and Implementation of an Object-Oriented Multimedia Knowledge Base Management System
028C/01 ŸdH.Ÿalshikawa
028C/02 ŸdF.ŸaSuzuki
028C/03 ŸdF.ŸaKozakura
028C/04 ŸdA.ŸaMakinouchi
028C/05 ŸdM.ŸaMiyagishima
028C/06 ŸdY.ŸaIzumida
028C/07 ŸdM.ŸaAoshima
028C/08 ŸdY.ŸaYamane
031A Ÿd18Ÿj1993Ÿe1Ÿh1-50Ÿg50
039B ŸcinŸ9129446335Ÿ8Association for Computing Machinery: ACM transactions on database systems. - New York, NY [u.a.] : ACM PressŸx19930000180001999
045W Ÿamat

Abb. 4.12 zu DDC-klassifizierender OLC-(Original)-Titeldatensatz aus Bsp. 4

In den nächsten beiden Beispielen geht es um erste Erkenntnisse bzgl. der Güte der automatischen Klassifikation, wobei mit dem Prototypen „COLIBRI-DDC“ ein Testlauf mit 50 (Bsp. 5) und einer mit 100 (Bsp. 6) Titeldatensätzen vorgenommen wird. In Bsp. 5 werden mit „0“ („Computer science, information, general works“) beginnende Titeldatensätze betrachtet. Mit „ul-ddc-pre ddc0-p.vim no > 50infile_to_ddc“ werden aus der Datei „ddc0-p.vim“[45] die ersten 50 Titeldatensätze präpariert, die nicht inhaltstragenden Wörter mit „ul-

[42] „file_to_ddc_olc_50328596x” aus Abb. 3.10, S. 14
[43] mit Ähnlichkeitswert „483“
[44] mit Ähnlichkeitswert „405“
[45] Der Name „ddc0-p.vim“ steht abkürzend für:
„ddc0“: GVK-Titeldatensatzmenge
„p“: internes Pica+ Präsentationsformat
„.vim“: Dateityp zum Öffnen mit Text-Editor „Vi Improved“

elim 50infile_to_ddc > 50infile_to_ddc_elim“ eliminiert und danach mit „ul-match-exact 50infile_to_ddc_elim“ (Variante 2) automatisch klassifiziert und für jede PPN mit den vom Fachpersonal intellektuell vergegebenen DDC-Notationen verglichen. Die ersten beiden zu klassifizierenden Titeldatensätze (PPN 337087628 und PPN 348332548) sind in Abb. 4.13 wiedergegeben:

```
edizioni
polistampa
1981-1998
catalogo
generale
DDC_NO= 015
PPN= 337087628
enterprise
javabeans
richard#monson-haefel
javabeans
java_(computer_program_language)
qa76.73.j38
DDC_NO= 005.2762
PPN= 348332548
…
```

Abb. 4.13 Auszug (die beiden ersten Titeldatensätze) aus „50infile_to_ddc_elim” (gesamte Datei im Anhang)

In 5:19 Stunden (Variante 1) bzw. 5:27 Stunden (Variante 2) wird folgendes Resultat erzielt (Bedeutung der Werte „+3“, „+2“, „+1“, „-„ in Abb. 4.15):

	Variante 1	Variante 2 (exact match)
+3	16%	68%
+2	8%	8%
+1	18%	4%
-	58%	20%

Abb. 4.14 Prozentuale Übereinstimmung zwischen intellektueller und automatischer DDC-Klassifizierung

Variante 1 erzielt eine 42%-ige Übereinstimmung, Variante 2 eine 80%-ige Übereinstimmung zwischen intellektueller und automatischer Klassifizierung bis zur 3. Stelle, wobei die Werte „+3“, ... „-„ für den Grad der Übereinstimmung zwischen intellektueller und automatischer DDC-Klassifizierung stehen sollen:

vollständige Übereinstimmung :	+3
min. erste 4 Stellen Übereinstimmung:	+2
min. erste 3 Stellen Übereinstimmung:	+1
keine Übereinstimmung:	-

Abb. 4.15 Werte für Übereinstimmungsgrade zwischen intellektueller und automatischer DDC-Klassifizierung, Übereinstimmungsgrade mehrerer Topkandidaten werden durch „/“ getrennt (z.B. Abb. 4.17, S. 24)

In Abb. 4.16 ist das Ergebnis der automatischen Klassifzierung der 50 Titeldatensätze (TTL 01-50) aus Bsp. 5 mit Übereinstimmungsgrad, PPN, intellektuell vergebener und automatisch ermittelter DDC-Notation mit Ähnlichkeitswert zusammengestellt. Im nächsten Bsp. 6 geht es um einen Testlauf einer (der DDC-Basis entnommenen) Stichprobe von 100 Titeldaten-sätzen, die per Zufallszahlengenerator[46] ausgewählt wurden. Am Ende des Anhangs befinden sich PPN-Bestimmung der 100 Titeldatensätze („PPN-Ermittlung von 100 Zufalls-PPN’s“) und deren Inhalt („100infile_to_ddc_elim “).

[46] In der Korn-Shell („ksh“) mit „echo $RANDOM“ erzeugt.

TTL	Ü	PPN	DDC_NO(intell.)	DDC_NO(automat.)	Ähnlichkeitswert
01	+3	119423278	016.823912	016.823912	5
02	+3	120269953	005.131	005.131	6
03	+3	184819415	070.509421	070.509421	6
04	+3	245133399	005.133	005.133	5
05	+2	302819258	005.44769	005.4469	8
06	+3	318112396	004.068	004.068	4
07	-	333112067	005.369	658.4/033	10
08	+3	333339568	025.0691	025.0691	8
09	-	333340760	004.35	005.2	8
10	-	334205891	005.13	670.42/7	5
11	-	334416833	005.369	519.5	10
12	+3	33456316X	005.133	005.13/3	8
13	+3	334951739	004.678	004.678	6
14	+3	335666507	001.94	001.94	12
15	+3	336227310	016.6863	016.6863	9
16	+3	336601743	005.1/13	005.1/13	6
17	+3	337087628	015	015	4
18	+3	337715122	005.2/82	005.2/82	5
19	+3	338111212	001.433	001.433	3
20	+3	338113347	020.941	020.941	10
21	+1	338113355	025.21	025.04	3
22	+3	33844520X	005.1/4	005.1/4	9
23	+3	338446214	003/.5	003/.5	8
24	+3	338501738	006.6869	006.6869	7
25	+3	338841709	005.72	005.72	10
26	+3	338980393	016.823/5	016.823/5	5
27	+3	339901217	005.1	005.1	8
28	+1	340116013	005.276	005.7/2	3
29	+2	340902817	005.71262	005.7/2	8
30	+3	341392723	070.44932490958	070.44932490958	10
31	+3	348332548	005.2762	005.2/762	3
32	+3	352029765	005.4469	005.4469	6
33	+3	353200379	025.0691	025.0691	12
34	+3	353559865	005.72	005.72	5
35	-	353559873	005.72	686.2/2544536	2
36	+3	355961482	005.72	005.72	7
37	+2	35618742X	005.74	005.7565	8
38	-	356188027	001.42	300.72	9
39	+3	356720578	070.50941	070.50941	13
40	+2	356730425	006.696	006.6869	6
41	+3	357090810	006.7	006.7	7
42	+3	357807324	004.678	004.678	9
43	+3	358380499	070.44	070.44	8
44	-	358710472	005.133	686.2/2544536	2
45	+3	358711185	005.1023	005.1023	5
46	-	358714648	004.68	942.1	3
47	+3	358897629	005.1068	005.1068	7
48	-	359527728	001.42	428.24	5
49	-	359706967	005.268	658.84	4
50	+3	359988911	004	004	4

Abb. 4.16 DDC-Klassifizierungsresultat von 50 Titeldatensätzen aus Hauptklasse „000“

Da Variante 2 in Bsp. 5 wesentlich besser abgeschnitten hat, wird in Bsp. 6 nur diese verwendet. In 13:13Stunden kommt das System zu folgendem Resultat (s. Abb. 4.17):

```
001 +3       017226643 519.5/028/55369 519.5/0285/5369                      4
002 +3       018741967 491.8 491.8                                         12
003 +3       018906559 745.6/19951 745.6'19951                              5
004 +3       025929364 581.5/2642/09435 581.5/2642/09435                   15
005 +3       031979645 011'.31 011'.31                                     12
006 +3       110124928 658.8/78/0685 658.8/78/0685                         13
007 +3       110787870 005.36/5 005.36                                      9
008 +3       111124220 519.5028553 519.5028553 10 300/.28/55369            10
009 +3       112179932 943 943                                              9
010 +3       112724280 354.54/792 354.54/792                                6
011 +3       113179707 155.9 155.9                                         18
012 +3       114469032 636.6/865 636.6/865                                  6
013 +3       11485369X 936.4 936.4                                          9
014 +3       114884994 658.1/6 658.1/6                                      7
015 +3       11574925X 831/.914099431 831/.914099431                       14
016 +3       115755802 344.30352044 344.30352044                           10
017 +3       115765719 973.921092 973.921/092 4 355.3/31/0924               4
018 -/-      115814752 248.3/4 {Meditation and contemplation}
                       299/.93 {Religions of eclectic and syncretistic
                                origin}                                      5
                       131.324 {Parapsychological and occult methods for
                                achieving well-being, happiness,
                                success--??}                                 5
019 +3       115877703 428.2/02465 428.2/02465                              8
020 +3       116173947 912/.1/3013294282 912/.1/3013294282                 18
021 +3       116296100 946/.108/092 946/.108/092                            7
022 -/+3     116376813 306/.0952 {Culture and institutions--
                                 Table 2. Geographic Areas, Historical
                                 Periods, Persons--Asia Orient
                                 Far East--Japan}
                       818/.5    {Jokes--American literature—
                                  1945-1999, ...}                            3
                       306/.0952 {Culture and institutions—
                                 Table 2. Geographic Areas, Historical
                                 Periods, Persons--Asia Orient
                                 Far East--Japan}                           3
023 +3       117321168 294.3/435/09593 294.3/435/09593                     15
024 +3       117432385 433/.1 433/.1                                        8
025 +3       117588881 786.6/3 786.6/3                                      9
026 +3       119437643 798.23 798.23                                        7
027 +3       119778424 821.914 821.914                                      5
028 +3       120640945 291.172 291.172                                     10
029 +3       122690214 230/.014 230/.014                                    8
030 +3       12396346X 070.4 070.4                                          6
031 +3       12587734X 005.4/469 005.4469                                   5
032 -/-/-/+3 127206507 016.91797/03 {Bibliographies and catalogs of
                                    works on specific subjects or in
                                    specific disciplines--Geography of
                                    and travel in North America--??}
                       973 {United States}                                  7
                       813/.54 {American fiction--1945-1999, ...}           7
                       610 {Medicine and health}                            7
                       016.91797/03 {see above}                             7
033 +3       132016559 553 552                                              4
034 +3       133253392 303.48/254061 303.48/254061                         14
035 +3       148035787 423/.1 423.1                                        33
036 +3       152498338 273/.9 273/.9                                        7
037 +3       153708158 658.15/26 658.15/26                                 10
038 +3       164906568 792 792                                              6
039 -/-      172809630 156 305.3 8 155.7                                    8
```

```
040 +3       181017040 822/.9/14 822/.9/14                                  6
041 +3       183173376 153.32 153.32                                        8
042 +3       185159710 838/.91209 838/.91209                               10
043 +3       186586701 721 721                                              4
044 +3       186923422 821/.2 821/.2                                        6
045 +3       18902481X 254/.5 254/.5                                       13
046 +3       192577352 144/.094 144/.094                                    9
047 +3       193983974 938/.05/072 938/.05/072                              4
048 +3       198344155 246/.7 246/.7                                       19
049 +3       213674947 813/.54 813/.54                                      3
050 -/-/+3   214678199 428.6   {Primers (Readers)--English language, ...}
                       823.914 {English fiction--1945-1999, ...}          3
                       741     {Arts & recreation--Arts--Arts &
                                recreation--Drawing and drawings}         3
                       428.6   {Primers (Readers)--English language, ...} 3
051 +3       218857659 959.803/5 959.803/5                                 12
052 +3       218881290 471/.1/09364 471/.1/09364                            4
053 +3       22509987X 536/.2 536/.2                                        9
054 +3       229347908 686.3/0092/2 686.3/0092/2                           16
055 +3       240511328 363.2/0954 363.2/0954                                6
056 +3       241640628 581.946/353 581.946/353                              8
057 +3       242349285 458.2421 458.2421                                    5
058 +2       24445521X 624.1/51 624.1/5                                     6
059 +3       245955569 005.2 005.2                                          5
060 +3       245964010 220/.046 220.046                                    12
061 +3       247146471 343.7305/2/08655 343.7305/2/08655                    8
062 +1       247562203 823.914 823                                          3
063 +3       249085550 422 422                                              8
064 +3       251047059 658.4/095 658.4/095                                 13
065 +3       252540832 782.421660922 782.421660922                          6
066 -/-/-/+3 267180756 150      {Psychology}
                       823/.912 {Literature--Literatures of specific
                                 languages and language families--English
                                 & Old English literatures—
                                 English fiction--20th century--
                                 Conrad, Joseph, ...}                      8
                       813/.52  {American fiction--1900-1945, ...}         8
                       616.89   {Mental disorders}                        8
                       150      {Psychology}                              8
067 +3       267540744 378.41 378.41                                        8
068 +3       268847363 423.09 423.09                                        8
069 +3       301122199 363.6/9/094977 363.6/9/094977                       11
070 +3       305854739 005.4469 005.4469                                    7
071 +3       312946104 507.1 507.1                                          6
072 +3       314795952 625.1/9 625.1/9                                     10
073 +3       315486899 914.531047 914.531047                               19
074 +3       315801123 599.5248 599.5248                                    3
075 +3       315826401 425 425                                              6
076 +3       31743246X 615.3210951 615.3210951                              6
077 +3       319071154 844/.3 844/.3                                        6
078 +3       321850963 150.19/5/092 150.19/5/092                            4
079 +3       322344697 823.914 823.914                                      4
080 +3       322391792 741.235 741.235                                      8
081 +3       323821278 338.951 338.951                                      6
082 +3       324539827 598/.09488 598/.09488                                6
083 +3       324967071 004/.35 004/.35                                     21
084 +3       327226048 796.8/15 796.8/15                                    8
085 +3       328600636 323.1/1965 323.1/1965                               16
086 +3       330141996 549 549                                              4
087 +3       332664821 025.06796352 025.06796352                            7
088 +3       334227267 813.54 813.54                                        5
089 -/+3     33454839X 730/.92/4 823.914 3 730/.92/4                         3
```

```
090 +3       336871074 220 220                                    3
091 +3       337172684 170 170                                    9
092 +3       33783718X 128 128                                   14
093 +3       338461914 647.94/068 647.94/068                     10
094 +3       346500230 941.11003 941.11003                       20
095 +3       347002684 973.046872910092 973.046872910092         12
096 +3       347919332 016.82154099729 016.82154099729           11
097 +3       350541183 751.4 751.4                                3
098 +3       352306343 193 193                                    8
099 +3       353780707 252.05233 252.05233                       23
100 +3       358663555 133.335 133.3/35                           5
```

Abb. 4.17 DDC-Klassifizierungsresultat von 100 per Zufall ausgewähltenTiteldatensätzen aus allen Hauptklassen

D.h. das System ermittelt im Bsp. 6 in 91% der Fälle exakt die DDC-Notation, die durch das DDC-Fachpersonal intellektuell vergeben wurde, in 93% der Fälle stimmen die ersten 3 Stellen überein:

	Variante 2 (exact match)
+3	91%
+2	1%
+1	1%
-	7%

Abb. 4.14 Prozentuale Übereinstimmung zw. intellektueller und automatischer DDC-Klassifizierung (Bsp. 6)

Resümee:

Die Ergebnisse der 6 Testbeispiele zeigen, daß in der vorliegenden Untersuchung ein aussichtsreicher Weg beschritten wird. Da Zufallsstichproben benutzt werden, weisen die Ergebnisse über die Beispiele hinaus und stützen die Hypothese, daß der erzielte Übereinstimmungsgrad zwischen automatischer und intellektueller Klassifikation auch für die Grundgesamtheit besteht. Diese Hypothese sowie weitere Testläufe und Effizienzverbesserungen der Programme sollen Gegenstand des nächsten VZG-Colibri-Berichtes 1/2004: „DDC-Basis: Grundlage der automatischen DDC-Klassifikation“ sein.