Automatische DDC-Klassifizierung bibliografischer Titeldatensätze der Deutschen Nationalbibliografie

Das Klassifizieren von Objekten (z. B. Fauna, Flora, Texte) ist ein Verfahren, das auf menschlicher Intelligenz basiert. In der Informatik – insbesondere auf dem Gebiet der Künstlichen Intelligenz (KI) – wird u. a. untersucht, inweit Verfahren, die menschliche Intelligenz benötigen, automatisiert werden können. Hierbei hat sich herausgestellt[1], dass die Lösung von Alltagsproblemen eine größere Herausforderung darstellt, als die Lösung von Spezialproblemen, wie z. B. das Erstellen eines Schachcomputers. So ist „Rybka" der seit Juni 2007 amtierende Computerschach-Weltmeister[2]. Inwieweit Alltagsprobleme mit Methoden der Künstlichen Intelligenz gelöst werden können, ist eine – für den allgemeinen Fall – noch offene Frage. Beim Lösen von Alltagsproblemen spielt die Verarbeitung der natürlichen Sprache, wie z. B. das Verstehen, eine wesentliche Rolle. Den „gesunden Menschenverstand" als Maschine (in der Cyc-Wissensbasis in Form von Fakten und Regeln) zu realisieren, ist Lenat's Ziel seit 1984. Bezüglich des KI-Paradeprojektes „Cyc" gibt es Cyc-Optimisten und Cyc-Pessimisten.[3]

Das Verstehen der natürlichen Sprache (z. B. Werktitel, Zusammenfassung, Vorwort, Inhalt) ist auch beim intellektuellen Klassifizieren von bibliografischen Titeldatensätzen oder Netzpublikationen notwendig, um diese Textobjekte korrekt klassifizieren zu können. Seit dem Jahr 2007 werden von der Deutschen Nationalbibliothek (DNB) nahezu alle Veröffentlichungen mit der Dewey Dezimalklassifikation (DDC) intellektuell klassifiziert.[4] Die Menge der zu klassifizierenden Veröffentlichungen steigt spätestens seit der Existenz des World Wide Web schneller an, als sie intellektuell sachlich erschlossen werden kann. Daher werden Verfahren gesucht, um die Klassifizierung von Textobjekten zu automatisieren oder die intellektuelle Klassifizierung zumindest zu unterstützen. Seit 1968[5] gibt es Verfahren zur automatischen Dokumentenklassifizierung (Information Retrieval, kurz: IR) und seit 1992 zur automatischen Textklassifizierung (Automated Text Categorization, kurz: ATC)[6]. Seit immer

---

[1] Scheitern des Projektes „General Problem Solver" von Herbert Simon und Allen Newell, 1957 – 1970. [http://de.wikipedia.org/wiki/General_Problem_Solver]; [http://www.zeit.de/zeit-wissen/2009/01/Titelstrecke-Frage7-Maschinen].

[2] Rybka: ca. 3.200 Elopunkte, Kasparov: ca. 2.851 Elopunkte. In: [Nüscheler 2009] Manfred Nüscheler: Rybka 3.0: Das beste PC-Schachprogramm der Welt. Netzathlethen GmbH, 5.5.2009. [http://www.netzathleten.de/Nachrichten/Rybka-3-0-Das-beste-PC-Schachprogramm-der-Welt/Schach/4285136507544480 34/a].

[3] [Schult 1996] Thomas J. Schult: Künstliche Intelligenz: Die Wissensbasis Cyc will Computern Selbstverständlichkeiten beibringen. Jetzt kann sie über das Internet genutzt werden. Die Zeit. 49/1996, 29.11.1996. [http://www.zeit.de/1996/49/cyc.txt.19961129.xml].

[4] [Scheven 2009] Esther Scheven: Inhaltserschliessung. DNB, Frankfurt am Main, 11.8.2009. [http://www.d-nb.de/wir/ueber_dnb/inhaltserschl.htm].

[5] [Salton 1968] Gerard Salton: Automatic Information Organization and Retrieval. McGraw-Hill, Inc., New York u.a., 1968, pp. 112, 133 — 135; [Bollmann/Konrad/Schneider/Zuse 1978] Peter Bollmann; Erhard Konrad; Hans-Jochen Schneider; Horst Zuse: Anwendung automatischer Klassifikationsverfahren mit dem System FAKYR. In: Wolfgang Dahlberg (Hrsg.): Kooperation in der Klassifikation. Proceedings der Sekt. 1 – 3 der 2. Fachtagung der Gesellschaft für Klassifikation e. V., Frankfurt-Höchst, 6. — 7. April 1978. Frankfurt, Gesellschaft für Klassifikation, 1978, S. 156 – 165. [http://www.gbv.de/dms/tib-ub-hannover/028342453.pdf].

[6] [Lewis 1992] David D. Lewis: An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, ACM Press, New York, US, 1992, pp. 37 – 50.

mehr digitale Objekte im World Wide Web zur Verfügung stehen, haben Arbeiten zur automatischen Textklassifizierung seit ca. 1998[7] verstärkt zugenommen. Dazu gehören seit 1996[8] auch Arbeiten zur automatischen DDC-Klassifizierung bzw. RVK-Klassifizierung[9] von bibliografischen Titeldatensätzen und Volltextdokumenten. Bei den Entwicklungen handelt es sich unseres Wissens bislang um experimentelle und keine im ständigen Betrieb[10] befindlichen Systeme. Auch das VZG[11]-Projekt Colibri/DDC[12] ist seit 2006 u. a. mit der automatischen DDC-Klassifizierung befasst. Die diesbezüglichen Untersuchungen und Entwicklungen dienen zur Beantwortung der Forschungsfrage: „Ist es möglich, eine inhaltlich stimmige DDC-Titelklassifikation aller GVK-PLUS[13]-Titeldatensätze automatisch zu erzielen?“

Colibri/DDC-Wettbewerb

Da in der Fachwelt starkes Interesse an (semi-)automatischen Klassifizierungssystemen von Textobjekten besteht und um den Anreiz für die Entwicklung solcher Systeme zu steigern, wurde auf dem 98. Deutschen Bibliothekartag in Erfurt der „Colibri/DDC-Wettbewerb“[14] initiiert, mit dem Ziel, den besten automatischen DDC-Klassifizierer für bibliografische Titeldatensätze zu finden. In Abbildung 1 ist ein fiktives Szenario für den Colibri/DDC-Wettbewerb skizziert: Während sich die Klassifizierungskomponente *vc_dcl* (*vzg colibri_ddc classifier*) des DDC-Suchsystems *vc_ds*[15] und das System Y schon länger in der Entwicklung befinden, gibt es ein System X, das kurz vor dem Start steht und ein System Z, das hiervon

---

[7] [Gabrilovich 2007] Evgeniy Gabrilovich: Bibliography on Automated Text Categorization – Bibliographic Statistics. Department of Computer Science, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel, October 2007. [http://liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html].

[8] [Oberhauser 2004] Otto Oberhauser: Automatisches Klassifizieren - Verfahren zur Erschliessung elektronischer Dokumente. Master's Thesis. Zusatzstudiengang Bibliotheks- und Informationswissenschaft, Fakultät für Informations- und Kommunikationswissenschaften, Fachhochschule Köln. Wien, 2004, S. 65.

[9] RVK: Regensburger Verbundklassifikation; [Pfeffer 2007] Magnus Pfeffer: Automatische Vergabe von RVK-Notationen anhand von bibliografischen Daten mittels fallbasiertem Schließen. Masterarbeit im Rahmen des postgradualen Fernstudiums Master of Arts. Humboldt-Universität zu Berlin. Philsophische Fakultät I. Institut für Bibliotheks- und Informationswissenschaft. Mannheim, Juni 2007. [http://blog.bib.uni-mannheim.de/Classification/wp-content/uploads/2007/10/main.pdf].

[10] Bis auf ev. rare Einzelfälle, wie z. B. „AutoDewey“, s. Fußnote 34.

[11] VZG: Verbundzentrale des Gemeinsamen Bibliotheksverbundes (GBV).

[12] Name und Ursprung des im Frühling 2003 aufgenommenen VZG-Projektes Colibri/DDC gehen auf das Pica-Projekt „Colibri (COntext generation and LInguistic tools for Bibliographic Retrieval Interfaces)“ zurück. [Pica, Jaarsverlag Annual report 97, ISSN 0168-992, ISBN 90-70311-87-9], S. 18.

[13] GVK-PLUS: Gemeinsamer Verbundkatalog (GVK) und Online Contents (OLC).

[14] [Reiner 2009b] Ulrike Reiner: Automatische DDC-Klassifizierung von bibliografischen Titeldatensätzen. 98. Deutscher Bibliothekartag: Ein neuer Blick auf Bibliotheken. TK10: Information erschließen und recherchieren. Blockveranstaltung: Inhalte erschließen – mit neuen Tools, Erfurt, 3.6.09. [http://www.opus-bayern.de/bib-info/volltexte/2009/736/]; Kurzversion: [http://www.slideshare.net/ulhumm/bibtag09-automat.ischeddcklassifizierunglite].

[15] [Reiner 2009b], S. 18 – 19 und S. 39. Systemnamen im Text in Kursivschrift.

noch etwas entfernt ist. Unabhängig vom Startzeitpunkt kann sich der Weg zum Erfolg unterschiedlich gestalten: Die Entwicklung der Systeme X und Y ist unbekannt; das zuletzt an den Start gegangene System Z könnte alle anderen Systeme „überholen", dann jedoch bei Verbesserungsversuchen in Stagnation geraten, ohne ans Ziel zu gelangen. *vc_dcl* könnte innerhalb der Entwicklung Rückschläge erfahren und überflüssige (Rück-)Wege beschreiten, dann jedoch (wie erhofft) ans Ziel gelangen. Ob das in Abbildung 1 illustrierte, fiktive Szenario des Colibri/DDC-Wettbewerbs in dieser oder ähnlicher Weise eintreten und ein einsatzbereiter, automatischer DDC-Klassifizierer Realität werden wird, wird die Zukunft zeigen.

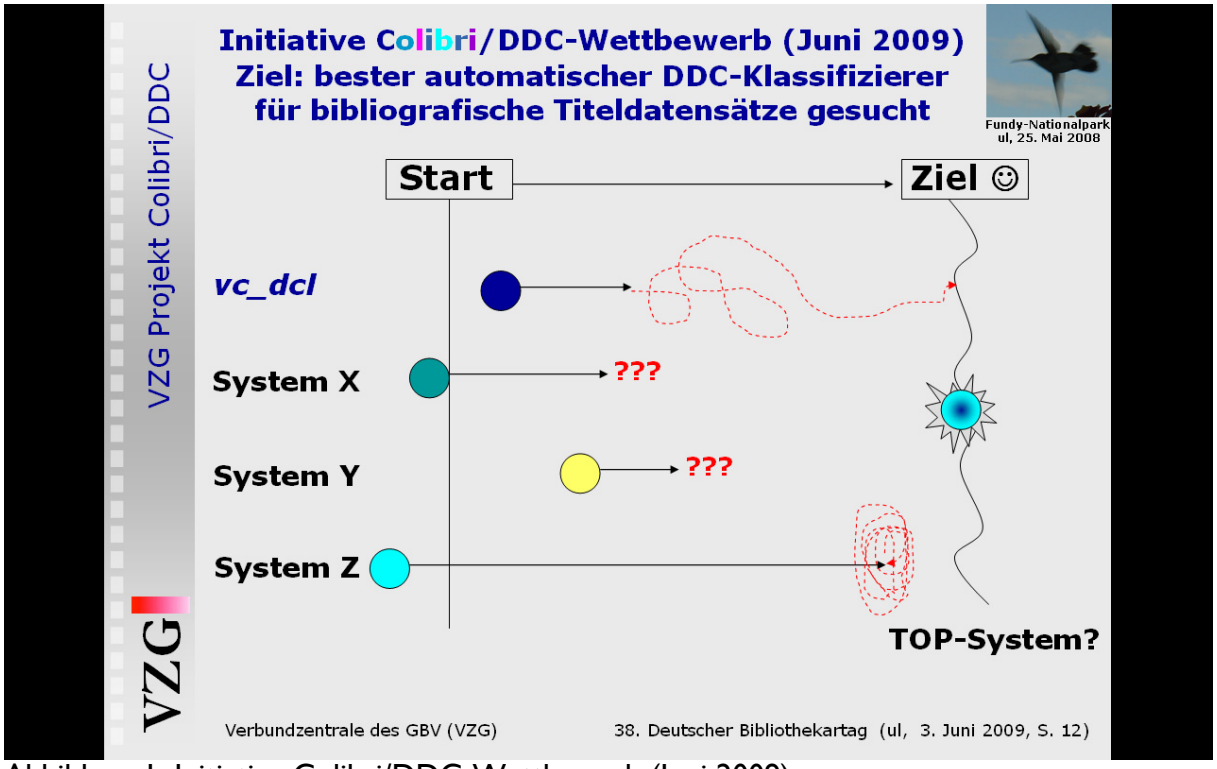

Abbildung 1: Initiative Colibri/DDC-Wettbewerb (Juni 2009)

Systemtests

Für einen Qualitätsvergleich von Systemen sind vergleichbare Testbedingungen unabdingbar. Diese sind zurzeit bei automatischen Klassifizierungssystemen noch nicht gegeben. Die Systeme differieren z. B. in Annahmen, Voraussetzungen, Anwendungsbereichen, zu klassifizierenden Objekten, Testbeständen und Bewertungsmaßen (ein Zustand, wie er 1983 im Information Retrieval vorherrschte[16]). Um für den „besten automatischen DDC-Klassifizierer" vergleichbare Testbedingungen zu erstellen, wurde zur Standardisierung der Colibri/DDC-Wettbewerb ins Leben gerufen. Nach [Salton/McGill 1983][17] sind für einen Systemtest mindestens folgende Bestandteile notwendig:

[16] [Reiner 1983] Ulrike Reiner: Experimente im Gebiet des Information Retrieval – Überblick und Stand der Forschung. Technische Universität Berlin, Institut für Angewandte Informatik, Fachbereich Informatik. LIVE-Bericht Nr. 6/83 (entstanden im Rahmen des vom BMFT geförderten Projektes „Leistungsbewertung von Information Retrieval Verfahren" (LIVE)), 1983. [http://portal.acm.org/citation.cfm?id=253495.253527].

[17] [Salton/McGill 1983] Gerard Salton; Michael J. McGill: Introduction to Modern Information Retrieval. Chapter 5: Retrieval Evaluation. McGraw-Hill International Book Company, Hamburg u.a. , 1983, S. 158.

1. Modell des Systems oder detaillierte Beschreibung des Systems und seiner Komponenten;
2. Zu testende Hypothesen;
3. Bewertungskriterien und Maße, die diese Kriterien widerspiegeln und
4. Methoden, die Daten zu ermitteln und zu bewerten.

Nach diesen vier Kriterien wird nachfolgend die Colibri/DDC-Klassifizierungskomponente *vc_dcl* genauer beschrieben und es werden zusätzlich einige andere in der Entwicklung befindlichen Systeme zur automatischen Klassifizierung zum Vergleich herangezogen.

1. Systeme zur automatischen Klassifizierung

Mit der Entwicklung der automatischen Klassifizierungskomponente *vc_dcl* des Suchsystems *vc_ds* innerhalb des Projektes Colibri/DDC[18] wurde im Jahr 2006 begonnen und sie wurde im Jahr 2008 fortgeführt[19]. Die klassifizierten Textobjekte sind bibliografische Titeldatensätze der Deutschen Nationalbibliografie[20]. Da die Titeldatensätze intellektuell vergebene DDC-Notationen enthalten, können daraus DDC-Testbestände[21] erstellt und (durch den Vergleich der intellektuell vergebenen mit den automatisch ermittelten DDC-Notationen) DDC-Klassifizierungssysteme bewertet werden (s. u. 3. Bewertungskriterien und -maße). Ein Aspekt des Colibri/DDC-Wettbewerbs ist – analog der Tradition des Information Retrieval (Cranfield, TREC, GOV2, CLEF, REUTERS u. a.[22]) – Standard-Testbestände für die automatische DDC-Klassifizierung aufzubauen / festzulegen. Bislang stehen die DNB-DDC-Testbestände *in_DNB-2007*, *in_DNB-2009* und *in_DNB-2009-2* im Colibri/DDC-Wettbewerb zur Verfügung, s. Tabelle 1[23]. Für die automatische Klassifizierung werden die MAB2-Dateien als Eingabedateien *in_DNB-2007*, *in_DNB-2009* und *in_DNB-2009-2* aufbereitet und als Menge von Objekt-Attribut-Wert-Tripeln (DDC-Notation, Deskriptor, Deskriptorwert)[24] repräsentiert. Inkorrekte DDC-Notationen und irrelevante Deskriptorwerte werden vorher

[18] [Reiner 2009a] Ulrike Reiner: VZG-Projekt Colibri. Bewertung von automatisch DDC-klassifizierten Titeldatensätzen der Deutschen Nationalbibliothek (DNB). VZG-Colibri-Bericht 1/2008, August 2008 – Februar 2009. [http://taipan.dyndns.org/~ul/colibri05.pdf].

[19] Details zu Colibri/DDC-Modell und Colibri/DDC-Suchsystem *vc_ds* mit seinen Komponenten s. [Reiner 2009a], S. 3-12 und [Reiner 2009b], S. 14 – 19; 23 – 26.

[20] In [Reiner 2009b], S. 6 ist z. B. der DNB-Titeldatensatz DNB984784829 im MAB2-Format wiedergegeben.

[21] Auf dem Gebiet des Information Retrieval / der Künstlichen Intelligenz auch Menge der Testdokumente bzw. Testkorpus genannt.

[22] [Reiner 2009b], S. 20.

[23] DNB: Deutsche Nationalbibliothek; *in* steht für input: *in_DNB-2007*: A0745-A0748, B0745-B0748, H0708-H0711; *in_DNB-2009*: A0912-A0915, B0912-B0915, H0812, H0901-H0903; *in_DNB-2009-2*: A0942-A0948, B0942-B0948, H0905-H0911; A, B, H stehen für Reihen der Deutschen Nationalbibliografie: Reihe A (Monografien und Periodika des Verlagsbuchhandels); Reihe B (Monografien und Periodika außerhalb des Verlagsbuchhandels); Reihe H (Hochschulschriften). Die nach „A", „B" und „H" folgenden ersten beiden Ziffern stehen für das Jahr, die letzten beiden Ziffern für die Kalenderwoche, also z. B. „A0745" für Reihe A aus der 45. KW des Jahres 2007. Weitere DDC-Testbestände für den Colibri/DDC-Wettbewerb sind erstrebenswert, z. B. DDC-Testbestände aus BASE- und NORBOK- bestehenden Titeldatensätzen, vgl. [Reiner 2009b], S. 21.

[24] DDC-Notation: Notation einer DDC-Klasse; Deskriptor: Pica+-Kategorie bzw. MAB2-Kategorie, deren Werte zur inhaltlichen Charakterisierung beitragen. Zurzeit berücksichtigte Pica+- bzw. MAB2-Deskriptoren s. [Reiner 2009b], S. 24 und 25; Deskriptorwert: Wert eines Deskriptors, s. [Reiner 2009b], S. 14.

eliminiert und relevante Deskriptorwerte u. a. transliteriert, deren Sonderzeichen entfernt und in Kleinschreibung transformiert.

DNB-DDC-Testbestände	Anzahl Titeldatensätze[25] alle	klassifizierte	Deskriptor-werte[26]	Sprache Deutsch	Englisch	andere	
in_DNB-2007	25.653	15.365	10,5	76%	20%	4%
in_DNB-2009	30.717	21.422	11,0	80%	18%	2%
in_DNB-2009-2	45.935	33.536	10,5	78%	20%	2%

Tabelle 1: Charakteristika dreier DNB-DDC-Testbestände

Die Klassifizierungsbasis für die DDC-Testbestände bildet *GVK-DDC*[27], die sich durch DDC-Lieferanten wie z. B. DNB, LoC und OCLC[28] jährlich vergrößert:

GVK-DDC	Anz. Titel-datensätze	*GVK-DDC0*	*GVK-DDC1*	*GVK-DDC2*	*GVK-DDC3*	*GVK-DDC4*	*GVK-DDC5*	*GVK-DDC6*	*GVK-DDC7*	*GVK-DDC8*	*GVK-DDC9*
2004	3,0 Mio.	99M	57M	109M	564M	46M	145M	306M	188M	324M	280M
2008	4,3 Mio.	236M	153M	253M	1.3G	118M	386M	677M	390M	666M	571M
2009	5,9 Mio.	515M	271M	394M	2.2G	358M	714M	1.2G	629M	981M	858M

Tabelle 2: GVK-DDC mit Teilmengen *GVK-DDC0* … *GVK-DDC9* (2004-2009)

Aus *GVK-DDC* wird die DDC-Datenbasis *vc_DB* (DataBase) und aus dem DDC-System[29] die DDC-Wissensbasis *vc_KB* (Knowledge Base) gebildet. *vc_DB* und *vc_KB* bilden zusammen *vc_DB_PLUS*, aus Effizienzgründen invertiert (*vc_IDB*: Inverted DataBase). Die Einlesezeit von *vc_IDB-2004* (510 MB) in den Hauptspeicher beträgt 3 Min., von *vc_IDB-2008* (712 MB) 5 Min. und von *vc_IDB-2009* (925 MB) 6 Min.[30] Die DNB-DDC-Testbestände *in_DNB-2007*, *in_DNB-2009* und *in_DNB-2009-2* werden größenabhängig in 2-3 Stunden automatisch klassifiziert. Der Klassifizierungsalgorithmus verwendet IR- und KI-Verfahren[31], jedoch bislang keine linguistischen Verfahren. Das Ergebnis der automatischen Klassifizierung wird intellektuell und automatisch bewertet. Weitere Klassifizierungsprojekte s. Tabelle 3.

---

[25] Die Differenz zwischen der Anzahl aller Titeldatensätze der Eingabedateien (*in_DNB-2007*, *in_DNB-2009* und *in_DNB_2009-2*) und der Anzahl der entsprechenden DDC-klassifizierten Titeldatensätze kommt dadurch zustande, da Titeldatensätze nur dann automatisch klassifiziert werden, wenn sie
1. eine korrekte DDCNotation enthalten,
2. in einem Klassifizierungslauf noch nicht klassifiziert wurden,
3. nicht bereits in der DDC-Daten-/Wissensbasis (hier: *vc_IDB-2008*, Erläuterung weiter unten im Text) enthalten sind.

[26] Durchschnittliche Anzahl Deskriptorwerte/Titeldatensatz.

[27] *GVK-DDC*: die Teilmenge des Gemeinsamen Verbundkataloges GVK, deren GVK-Titeldatensätze mindestens eine DDC-Notation enthalten, s. [Reiner 2009a], S. 4; *GVK-DDC0*: Titeldatensätze, die mit „0“ beginnende DDC-Notationen enthalten, … , *GVK-DDC9*: Titeldatensätze, die mit „9“ beginnende DDC-Notationen enthalten. *GVK-DDC-2009* wurde im Dezember 2009 erstellt.

[28] DNB: Dt. Nationalbibliothek; LoC: Library of Congress; OCLC: OCLC Online Computer Library Center.

[29] Das DDC-System liegt Colibri/DDC in elektronischer XML-Form in der 22. englischsprachigen Auflage vor.

[30] Verwendete Hard- und Software: HP Proliant DL585 G1, 4xAMD Opteron 275, 2.2 GHz, 16 GB Hauptspeicher. SuSE Linux Enterprise 10, gawk-3.1.5.

[31] [Reiner 2009a], S. 4 – 13 und [Reiner 2009b]. S. 15-17.

2. Hypothesen

Einzelne Hypothesen, Forschungsfragen, Annahmen etc. werden hier exemplarisch wiedergegeben (Quellenangaben s. Fußnote 34).

Klassifizierungsprojekt	**Auto-DDC**	**Auto Dewey**	**Colibri/ DDC**	**Pfeffer/ RVK**	**Scorpion/ DDC**	**TopicModels/ DDC**
Laufzeit	seit 2006	seit 2006	seit 2006/2008	seit 2005	1996-2000	seit 2009
Klassifizierung	DDC	DDC	DDC	RVK	DDC	DDC
zu klassifizierende Textobjekte	bibliograf. Titeldatensätze	bibliograf. Titeldatensätze	bibliograf. Titeldatensätze	bibliograf. Titeldatensätze	Elektron. Webdokumente (Volltexte)	bibliograf. Titeldatensätze
berücksichtigte Daten der zu klassifizierenden Textobjekten	inhaltsbezogene Kategorien (MARC21-Titeldatensätze)	engl., franz., ital. literar. Autoren: Dichtung, Dramen, Prosa	Deskriptorwerte von inhaltstragenden Deskriptoren (MAB2-Kategorien)	Titel-/ Schlagwörter	„Editorial Support System (ESS)"-Datensätze des DDC-Systems	OAI Metadaten: Titel-/Schlagwörter, Zusammenfassungen
Testbestände	22.110 LoC-Titeldatensätze aus Science & Technology (BDS&T)	unbekannt	*in-DNB-2007* *in-DNB-2009* *in-DNB-2009-2*	10.000 zufällige Titeldatensätze aus Fallbasis	NetFirst ([Shafer 1997], s. Fußnote 34)	719 englische Titeldatensätze 1.000 deutsche Titeldatensätze
Basis für die Klassifzierung	66.440 LoC-Titeldatensätze aus Science & Technology (BDS&T)	MARC21-Titeldatensätze	GVK-DDC-Titeldaten-Sätze & DDC-System (*vc_IDB*)	SWB-/ HeBIS-Titeldatensätze (Fallbasis)	DDC-System als Wissensbank (Scorpion Dewey Database)	100.000 BASE[32]-Dokumente, Wikipedia (Trainingsdokumente)
Algorithmen aus dem/den Gebiet/en	KI-ATC: Naïve Bayes, Support Vector Machines, DDC restructuring	unbekannt (Menge von Algorithmen)	IR: Ähnlichkeitsmaß & KI: heuristische Funktion	KI: fallbasiertes Schliessen	IR: Ähnlichkeitsmaße (SMART[33])	KI-ATC: Support Vector Machines, Latent Semantic Analysis, SEQ-based classifier, Wikipedia-based classifier
Hypothesen / Fragen	ja	unbekannt	Ja	nein	ja	ja
Bewertung	automatisch	unbekannt	intellektuell u. automatisch	intellekt. u. automatisch	automatisch	automatisch
Automatisierungsgrad	semi-automatisch	(semi-) automatisch	automatisch	automatisch	automatisch	automatisch

Tabelle 3: Klassifizierungsprojekte als potenzielle Kandidaten im Colibri/DDC-Wettbewerb[34]

[32] [Pieper/Summann 2006] Dirk Pieper; Friedrich Summann: Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service. Library Hi Tech, Bd. 24, Nr. 4, 2006, pp. 614-619. [http://eprints.rclis.org/9160/1/pieper_summann_final_web.pdf].

[33] [http://en.wikipedia.org/wiki/SMART_Information_Retrieval_System].

[34] Auto-DDC [Wang 2009] Jun Wang: An Extensive Study on Automated Dewey Decimal Classification. Journal of the American Society for Information Science and Technology, Vol. 60, No. 11, 2009, pp. 2269 – 2286.

Auto-DDC (Forschungsfragen, S. 2270): 1. Sind bibliothekarische Klassifikationsschemata für eine maschinelle Verarbeitung geeignet? 2. Sind Methoden des überwachten Lernens geeignet, um effektiv auf bibliografischen Daten zu arbeiten?

Colibri/DDC (Hypothesen, [Reiner 2009b], S. 41-47): Es gibt signifikante Unterschiede zwischen 1. unterschiedlichen DDC-Daten-/Wissensbasen, 2. den DDC-Klassen, 3. den DNB-Reihen A, B und H und 4. hinsichtlich der Stelligkeit der DDC-Notationen. Es gibt keine signifikanten Unterschiede zwischen 5. unterschiedlichen Testbeständen und 6. englischen und deutschen DNB-Titeldatensätzen.

Pfeffer/RVK (Testbeschränkungen, S. 8): keine Reihen, keine Zeitschriften, keine formalen Klassifikationen.

Scorpion/DDC (Forschungsfragen[35]): 1. Welchen Effekt hätte die zusätzliche Verwendung von „Library of Congress Subject Headings" (LCSH)? 2. In welcher Weise änderten sich die Ergebnisse, wenn andere Algorithmen als ATN/ATC verwendet würden? 3. Welches Ergebnis lieferten andere Klassifikationen, wie z. B. die „Library of Congress Classification" unter ähnlichen Bedingungen?

TopicModels/DDC (Hypothese, S. 7): Der Inhalt eines wissenschaflichen Dokumentes wird mithilfe seines Titels, Schlagwörtern und einer kurzen Beschreibung zuverlässig klassifiziert.

AutoDewey [Tillett 2008] Barbara B. Tillett: Library of Congress Report. ALA ALCTS Committee on Cataloging: Description and Access. Midwinter Meeting, Philadelphia, PA, January 12, 2008. [http://www.libraries.psu.edu/tas/jca/ccda/docs/lc0801.pdf]; [Green 2008] Rebecca Green: Literary Authors: AutoDewey and LC Name Authority File. Dewey Breakfast/Update. ALA Midwinter Meeting, January 12, OCLC, 2008. [http://www.oclc.org/dewey/discussion/papers/literary_authors.ppt].

Colibri/DDC [Reiner 2009a]; [Reiner 2009b].

Pfeffer/RVK [Pfeffer 2008] Magnus Pfeffer: Automatische Vergabe von RVK-Notationen mittels fallbasiertem Schließen. 97. Deutscher Bibliothekartag, 5. Juni 2008, Mannheim. [http://blog.bib.uni-mannheim.de/Classification/wp-content/uploads/2008/6/bibtag2008_rvk.pdf].

Scorpion/DDC [Shafer 1996] Keith Shafer: A Brief Introduction to Scorpion. OCLC, Dublin, Ohio, 1996. [http://www.worldcat.org/oclc/54388055]; [Subramanian/ Shafer 1997] Srividhya Subramanian; Keith Shafer: Clustering. OCLC, Dublin, Ohio, 1997. [http://www.worldcat.org/oclc/54084278]; [Shafer 1997] Keith E. Shafer: Evaluating Scorpion Results. Annual Review of OCLC Research, 1997. [http://worldcat.org/arcviewer/1/OCC/2003/06/03/000000341 1/viewer/file1.html]. [Shafer/Subramanian/Fausey 1998] Keith Shafer; Srividhya Subramanian; Jon Fausey: Measures for Evaluating Automatic Subject Assignment of Electronic Resources. OCLC, Dublin, Ohio, USA, 1998. [http://www.worldcat.org/oclc/54084501].

TopicModels/DDC [Mehler/Waltinger 2009] Alexander Mehler; Ulli Waltinger: Enhancing Document Modeling by Means of Open Topic Models: Crossing the Frontier of Classification Schemes in Digital Libraries by Example of the DDC. Library Hi Tech, Vol. 27, Issue 4, 2009, pp. 520 – 539.

[35] S. 43 in [Thompson/Shafer/Vizine-Goetz 1997] Roger Thompson; Keith Shafer; Diane Vizine-Goetz: Evaluating Dewey Concepts as a Knowledge Base for Automatic Subject Assignment. International Conference on Digital Libraries Archive. Proceedings of the Second ACM International Conference on Digital Libraries. Philadelphia, Pennsylvania, United States, 1997, pp. 37-46. [http://elvis.slis.indiana.edu/irpub/DL/1997/pdf5.pdf].

3. Bewertungskriterien und -maße

Um die Qualität der Klassifizierungsergebnisse beurteilen zu können, müssen diese bewertet und wegen der Vergleichbarkeit müssen diesselben Bewertungskriterien und -maße verwendet werden. Dies ist zurzeit bei der automatischen Klassifizierung von bibliografischen Titeldatensätzen noch nicht der Fall. Bei den Klassifizierungsprojekten sind verschiedene Bewertungsmaße im Einsatz. Ausserdem kann es sein, dass Berechnungen mit demselben Maß unterschiedlich vorgenommen werden (Micro-/Macroaverage von Precision/Recall[36]). Im Projekt Colibri/DDC ist durch die DNB eine intellektuelle (exakte, gute, mittlere, keine, ausreichende Übereinstimmung, Hauptsachgruppentreffer) und durch die *vc_ds*-Programm-komponente *vc_dce*[37] eine automatische Bewertung (u. a. mit den Bewertungsmaßen *C, CP* und *CN*) und anhand von Beispielen eine vergleichende Betrachtung von Bewertungen der Klassifizierungsergebnisse vorgenommen worden[38]. Andere Projekte verwenden: Auto-DDC: Macro-/(depth-based)Microaverage von Precision und Recall; Pfeffer/RVK: perfekte, gute, mäßige und schlechte Übereinstimmung; Scorpion/DDC: Relationen: übereinstimmend, allgemeiner als, korreliert, synonym, exakt und thematisch nah; TopicModels/DDC: Precision, Recall und F-Score.

4. Tests und Bewertung

In den o. g. Projekten sind einige Tests mit unterschiedlichen Bewertungen durchgeführt worden. Diese Arbeit soll die notwendige Reproduzierbarkeit der Tests und die Vergleichbarkeit unterschiedlicher Klassifizierungsergebnisse fördern, um den besten automatischen (DDC-)Klassifizierer für bibliografische Titeldatensätze ermitteln zu können.

Ergebnisse und Ausblick

Für den Stand der Entwicklungen seien einzelne Ergebnisse / Einschätzungen wiedergegeben:

Scorpion/DDC: "Scorpion cannot replace human cataloging. There are many aspects of human cataloging that are difficult if not impossible to automate." [Shafer 1996]

LCC: "Fully automatic classification might not be possible considering the size and diversity of the LCC scheme" [Larson 1992[39]].

---

[36] [http://en.wikipedia.org/wiki/Precision_and_recall]; [Lewis 1991] David D. Lewis: Evaluating Text Categorization. Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics. Morristown, NJ, USA, 1991, pp. 312 – 318. [http://portal.acm.org/citation.cfm?id=112471].

[37] *vc_dce*: *vzg colibri_ddc classification results evaluator* [Reiner 2009b], S. 39.

[38] [Reiner 2009a], S. 13 ff; [Reiner 2009b], S. 31ff.

[39] [Larson 1992] Ray R. Larson: Experiments in Automatic Library of Congress Classification. Journal of the American Society for Information Science, Vol. 43, pp. 130 – 148; zitiert nach [Wang 2009], S. 2281.

Auto-DDC: „With no more than three {user}[40] interactions, a classification accuracy of nearly 90 % is achieved, thus providing a practical solution to the automatic bibliographic classification problem“ [Wang 2009], S. 2269.

Pfeffer/RVK: Die SWB-Datenbasis mit 2.496.839 Titeln erzielte 57,26 % (Hamming) und 56,89 % (IDF) „perfekte“ und 18,99 % (Hamming) und 18,84 % (IDF) „gute“ Ergebnisse [Pfeffer 2008], S. 13; die intellektuelle Überprüfung (SWB, HeBIS) kommt jedoch zu dem Schluss, „dass die Qualität der automatisch generierten Klassifikationen zu schlecht ist, um direkt in die Verbunddatenbanken eingespielt zu werden.“[41]

TopicModels/DDC: „By this procedure we get a classification value for each main class of the DDC which expresses the relatedness of a given OAI-input stream to the selected class ... With an overall F-score of .761, SVMs provide an adequate DDC-related method to classify documents based on their OAI metadata.” [Mehler/Waltinger 2009], S. 8/14.

Colibri/DDC: In 63.85 % (*CN*-Wert) stimmen intellektuell vergebene und automatisch ermittelte DDC-Notationen in der DDC-Hauptklasse überein (bei *vc_DB-2008* und *in_DNB-2009* unter genannten Voraussetzungen in [Reiner 2009a], Fußnoten 45 und 52). Weitere Ergebnisse in [Reiner 2009b], S. 41 – 47. Durch Eliminierung weiterer Banalwörter erhöht sich der *CN*-Wert um ca. 1 %. Der neueste Test-bestand *in_DNB-2009-2* liefert mit der neuesten Klassifizierungsbasis *GVK-DDC-2009* einen *CN*-Wert von ca. 91 %! Dieses erstaunliche Ergebnis muss näher untersucht werden (von insges. 45.935 werden nur 13.115 Titeldatensätze klassifiziert, s. Fußnote 25). Mit hoher Wahrscheinlichkeit ist eine große Menge weiterer Titeldatensätze des DNB-Testbestandes in der neuen *GVK-DDC-2009* enthalten, ohne mit den derzeitigen DNB-Kennungs-, ISBN- und ISSN-Prüfungen erkannt worden zu sein. Wenn dies der Fall ist, müssen weitere Titeldatensatz-Identifizierungs-Prüfungen (auf Enthaltensein in der Klassifizierungsbasis), z. B. auf Titel-Person-Basis, implementiert werden.

Fazit

Die inspizierten automatischen (DDC-)Klassifizierer arbeiten bereits besser als der Zufall, aber für einen professionellen umfangreichen Einsatz für Millionen von zu klassifizierenden (z. B. GVK-)Titeldatensätzen sind sie noch nicht geeignet. Auto-DDC erreicht zwar 90 % „accuracy“, dieses Ergebnis wird jedoch nur in Kombination mit der intellektuellen Leistung (bis zu drei Benutzeraktionen) erreicht. In den nächsten Schritten müssen die Klassifizierungsergebnisse verbessert und wegen der angestrebten Vergleichbarkeit im Colibri/DDC-Wettbewerb diesselben Hypothesen geprüft und die Ergebnisse in gleicher Weise bewertet werden.

[40] {user}: Einfügung durch Autorin.

[41] „Zurück auf Los“ [http://blog.bib.uni-mannheim.de/Classification/?p=30].

Dank

Die Autorin dankt Kristina Knull-Schlomann (DNB) für ihre Ermunterung, in „Dialog mit Bibliotheken“[42] eine schriftliche Version des von der Autorin auf dem 98. Deutschen Bibliothekartag in Erfurt gehaltenen Vortrags in aktualisierter Form zu veröffentlichen und zudem den DNB- und VZG-KollegInnen für die zur Verfügungstellung der Daten zur Erstellung der DNB-DDC-Testbestände und GVK-DDC-Klassifizierungsbasen, insbesondere (in alphabetischer Reihenfolge): Angelika Cremer-Reiber (DNB), Reiner Diedrichs (VZG), Siegfried Kalb (VZG) und Claudia Werner (DNB).

[42] [http://www.d-nb.de/service/publikationen/dialog_10_01.htm]. Die hier veröffentlichte Version [http://taipan.dyndns.org/~ul/dialog2010.pdf] enthält umfangreichere Erläuterungen und Quellenangaben.