Automatisches Klassifizieren und Bibliothekskataloge

Otto Oberhauser

Übersicht

1. Einleitung
2. Ausgangsdaten
3. Automatisches Indexieren und Bibliothekskataloge
4. Automatisches Klassifizieren
   4.1 Methodische Vorbemerkungen
   4.2 Die LCC-Studie von LARSON
   4.3 Weitere Untersuchungen
5. Fazit: Automatisches Klassifizieren und Bibliothekskataloge

1. Einleitung

In der bibliothekarischen Welt sind die Vorzüge einer *klassifikatorischen* Inhaltserschließung seit jeher wohlbekannt. Auch im Zeitalter der Online-Kataloge gibt es dafür keinen wirklichen Ersatz, da – kurz formuliert – ein stichwortbasiertes Retrieval alleine mit Problemen wie Ambiguität und Mehrsprachigkeit nicht fertig zu werden vermag. Zahlreiche Online-Kataloge weisen daher Notationen verschiedener Klassifikationssysteme auf; allerdings sind die darauf basierenden Abfragemöglichkeiten meist noch arg unterentwickelt. Viele Datensätze in OPACs sind aber *überhaupt nicht* sachlich erschlossen, sei es, dass sie aus retrospektiv konvertierten Nominalkatalogen stammen, sei es, dass ein Mangel an personellen Ressourcen ihre inhaltliche Erschließung verhindert hat. Angesichts großer Mengen solcher Datensätze[1] liegt ein Interesse an automatischen Verfahren zur Sacherschließung durchaus nahe.

[1] Im Österreichischen Verbundkatalog beträgt der Anteil der Datensätze mit verbaler Sacherschließung nur 44 % (April 2005), was aber im Vergleich mit anderen deutschsprachigen Verbünden sogar als guter Wert gelten kann.

2. Ausgangsdaten

Dieser Beitrag beschäftigt sich mit den Perspektiven des Einsatzes von Techniken des *automatischen Klassifizierens* in Bibliotheken. Dabei geht es um Ansätze, die darauf abzielen, mit Hilfe eines geeigneten Algorithmus die zu erschließenden Dokumente mit Notationen eines vorgegebenen Klassifikationssystems zu versehen – im Gegensatz zum so genannten *Clustern,* das auch die automatische Identifizierung der klassifikatorischen Struktur einer Dokumentenkollektion miteinschließt. Letzteres ist für Bibliotheken wenig interessant, da die klassifikatorische Inhaltserschließung in der Regel auf der kontinuierlichen Verwendung eines *bestehenden* Schemas – von einfacheren „Haussystematiken" bis hin zu komplexen Systemen wie z.B. der Dewey Dezimalklassifikation (DDC) – beruht.

Wie auch im Fall des *automatischen Indexierens* sind für das automatische Klassifizieren in Online-Katalogen *zwei* typische Anwendungsfälle denkbar. Im einen – derzeit noch wesentlich häufigeren – Fall liegen die zu klassifizierenden Dokumente in Form normaler OPAC-Datensätze vor, d.h. als Katalogisate, die sich aus den Daten der Formalerschließung und (zum Teil) der verbalen Sacherschließung zusammensetzen. Im zweiten Fall verfügen diese Datensätze daneben auch noch über Anreicherungen in Form von Inhaltsverzeichnissen, Abstracts oder gar Volltexten, wodurch ein wesentlich umfangreicheres Vokabular für darauf aufsetzende Techniken der automatischen Indexierung oder Klassifizierung zur Verfügung steht. Dieser Beitrag beschäftigt sich mit dem erstgenannten Fall, also mit jener großen Menge von Datensätzen, für die nur Angaben wie Autor, Titel, Untertitel, Impressum und eventuell Schlagwörter bzw. Schlagwortketten vorliegen.

3. Automatisches Indexieren und Bibliothekskataloge

Vorweg ein kurzer Überblick zum automatischen *Indexieren:* Dabei werden Verfahren eingesetzt, „die vollautomatisch Dokumente analysieren und abgeleitet aus dieser Analyse entweder ausgewählte Terme aus dem Dokument extrahieren und – unter bestimmten Verfahrensvoraussetzungen in einer bearbeiteten Form – als Indexterme abspeichern (Extraktionsverfahren) oder Deskriptoren einer kontrollierten Indexierungssprache dem Dokument als Inhaltsrepräsentanten zuweisen (Additionsverfahren)".[2] Für OPACs, in denen die Mehrzahl der Katalogisate in deutscher Sprache vorliegt und bei denen sich das inhaltstragende Wortmaterial in der Regel auf Titel- bzw. Schlagwörter beschränkt, eignen sich primär *linguistische* Indexierungsverfahren, die mit verschiedenen Methoden – Reduktion auf Grundformen, Zerlegung von Komposita usw. – eine Standardisierung bzw. Normierung dieses

[2] *Nohr*, H. (2003). *Grundlagen der automatischen Indexierung: Ein Lehrbuch*. Berlin: Logos-Verl. (hier: S. 20).

Vokabulars erzielen. Im Englischen eignen sich hiefür relativ einfache, regelbasierte Techniken. Für deutschsprachige Texte mit ihren abweichenden Mehrzahl- und Flexionsformen, Kompositabildungen und weiteren Unregelmäßigkeiten werden dagegen Verfahren benötigt, die auf sehr umfangreichen Wörterbüchern beruhen, in denen die konkret auftretenden Wortformen und ihre gewünschte Umsetzung mehr oder weniger detailliert lexikalisiert sind.

Schon in den 1990er Jahren wurden an der ULB Düsseldorf zwei Projekte durchgeführt, im Rahmen derer die Eignung eines wörterbuchbasierten Verfahrens für deutsche OPACs getestet wurde. Während im Projekt *MILOS I* (1993–1995) 40.000 Titel aus dem OPAC der ULB Düsseldorf maschinell indexiert und einem Retrievaltest mit 50 Suchfragen unterzogen wurden,[3] basierte das *MILOS II*-Projekt (1995–1996) auf der Indexierung von rund 190.000 Titelsätzen der Deutschen Nationalbibliographie, der Einbindung der SWD in das Indexierungsverfahren sowie auf einem Retrievaltest mit 100 Suchfragen in einer Allegro-Umgebung.[4] Ein weiterer Test mit ca. 47.000 ekz-Datensätzen und 30 Suchfragen wurde im Rahmen einer bibliothekarischen Diplomarbeit in Bonn durchgeführt.[5] Alle drei Studien gelangten zu dem Schluss, dass durch das automatische Indexieren eine beträchtliche Steigerung des Recalls bzw. der Menge relevanter Treffer erzielt wurde, ohne dass es zu dramatischen Einbußen hinsichtlich der Präzision kam. Außerdem ging dadurch die Zahl der Nulltreffer-Resultate – das Problem vieler OPAC-Benutzer – deutlich zurück. Diese Ergebnisse wurden kürzlich auch in einem weiteren Test mit über 70.000 zufällig ausgewählten Datensätzen aus dem Österreichischen Verbundkatalog – in einer Aleph-500-Umgebung – bestätigt.[6]

Diesen Erfolg versprechenden Befunden zum Trotz ist der praktische Einsatz des automatischen Indexierens bislang auf wenige deutschsprachige Bibliotheken beschränkt geblieben. Die Gründe für diese Zurückhaltung mögen

[3] *Lepsky*, K. (1994). *Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen*. Köln: Greven. (Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen; 18). – DERS. (1996). Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. In: NIGGEMANN, E.; *Lepsky*, K. (Hrsg.) *Zukunft der Sacherschließung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995*. Düsseldorf: Universitäts- und Landesbibliothek. S. 12–36. – LEPSKY, K.; SIEPMANN, J.; ZIMMERMANN, H.: Automatische Indexierung für Online-Kataloge: Ergebnisse eines Retrievaltests. *Zeitschrift für Bibliothekswesen und Bibliographie.* 43(1). S. 47–56.

[4] *Gödert*, W.; *Liebig*, M. (1997). Maschinelle Indexierung auf dem Prüfstand: Ergebnisse eines Retrievaltests zum MILOS II Projekt. *Bibliotheksdienst*. 31(1). S. 59–68. – SACHSE, E.; LIEBIG, M.; GÖDERT, W. (1998). *Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt*. Köln: FH Köln. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; 14).

[5] *Grummann*, M. (2000). Sind Verfahren zur maschinellen Indexierung für Literaturbestände Öffentlicher Bibliotheken geeignet? Retrievaltests von indexierten ekz-Daten mit der Software IDX. *Bibliothek: Forschung und Praxis*. 24(3). S. 297–318.

[6] *Oberhauser, O.; Labner,* J. (2003). OPAC-Erweiterung durch automatische Indexierung: Empirische Untersuchung mit Daten aus dem Österreichischen Verbundkatalog. *ABI-Technik*. 23(4). S. 305–314.

in einer zögerlichen Einstellung der Praktiker gegenüber automatischen Verfahren und in der fehlenden Rezeption informationswissenschaftlicher Ergebnisse durch die Hersteller von Bibliothekssoftware liegen, nicht zuletzt aber auch im geringen Interesse bibliothekarischer Entscheidungsträger für Fragen der inhaltlichen Erschließung.

4. Automatisches Klassifizieren

Wie umfassende Recherchen des Verfassers zeigten, gibt es im Gegensatz zum automatischen Indexieren bislang *keine einzige* Untersuchung, die sich mit dem automatischen Klassifizieren deutschsprachiger Katalogisate beschäftigt. Auch für den fremdsprachigen Bereich konnten nur wenige Studien gefunden werden. Diese sollen im Folgenden vorgestellt werden.

4.1 Methodische Vorbemerkungen

Auf die Methodik des automatischen Klassifizierens kann hier nur kursorisch eingegangen werden. Für eine ausführlichere Darstellung sei auf die jüngst publizierte Einführung des Verfassers[7] oder etwa die Beiträge von *Sebastiani*[8] verwiesen. In Kürze sei jedoch erwähnt, dass sich seit den 1990er Jahren, zunächst in der Forschung und inzwischen auch schon im Umfeld kommerzieller Software, für das automatische Klassifizieren von Textdokumenten der Einsatz maschineller Lernverfahren durchgesetzt hat. Darauf basierende Systeme bestehen stets aus zwei Komponenten:

- Einer Komponente zum *Wissenserwerb* in der *Trainingsphase;* dabei werden auf der Grundlage einer Menge bereits intellektuell klassifizierter *Trainingsdokumente* die Charakteristika der Klassen gelernt und Klassenprofile erstellt;
- der eigentlichen Komponente zum *Klassifizieren* („Klassifikator") in der darauf folgenden *Klassifizierungsphase*, in der neue, d.h. noch nicht klassifizierte Dokumente hinsichtlich ihrer Charakteristika analysiert und durch einen Vergleich mit den Klassenprofilen den passenden Klassen zugeordnet werden.

Für die Evaluierung der durch das automatische Verfahren erzielten Ergebnisse wird vor der Erstellung des Klassifikators die Trainingsmenge in zwei Teile geteilt:

---

[7] *Oberhauser*, O. (2005). *Automatisches Klassifizieren: Entwicklungsstand – Methodik – Anwendungsbereiche*. Frankfurt/M. usw., Peter Lang. (Europäische Hochschulschriften, Reihe XLI, Informatik; 43). [hier: S. 17–38].

[8] *Sebastiani*, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*. 34(1). S. 1–47. – *Ders.* (2005). Text categorization. In: *Text Mining and Its Applications*. Hrsg.: *Zanasi*, A. Southampton, WIT Press. (Management information systems; 3). S. 109–129.
Preprint: <http://www.math.unipd.it/~fabseb60/Publications/TM05.pdf> [26. 06. 2005]

- Die *Trainingsdokumente*, auf deren Basis der Klassifikator gebildet wird;
- die *Testdokumente*, die für den Test auf Güte herangezogen werden, wobei jedes Dokument automatisch klassifiziert und das Resultat mit der intellektuellen Klassifizierung verglichen wird.

Damit ein Klassifikator die zu verwendenden Texte interpretieren kann, wird ein Indexierungsverfahren benötigt, das eine kompakte Repräsentation dieser Texte zum Ergebnis hat. Üblicherweise geschieht dies durch die Bildung eines *Vektors von Termgewichten*, der ausdrücken soll, in welchem Ausmaß jeder im Dokument auftretende Term („Attribut", „Merkmal") zur Bedeutung des betreffenden Trainings- bzw. Testdokuments beiträgt. Dabei gibt es verschiedene Möglichkeiten für die Definition eines Attributs und die Berechnung der Termgewichte. Vor der Indexierung wird meist eine *Textnormalisierung* durchgeführt, bei der alle unerwünschten Zeichen bzw. Terme entfernt werden. Dabei gelangen häufig auch *linguistische Verfahren* zur Anwendung, insbesondere die Eliminierung von Stoppwörtern und die Lemmatisierung bzw. Stammformenbildung. Bei deutschsprachigen Texten sollten zudem Kompositazerlegung, Derivation (Zusammenfassen von verschiedenen Wortklassen oder Derivaten, wie z.B. Adjektiven, Substantiven und Verben mit derselben Grundform) und Bindestrichergänzung erfolgen. Da die bei Verwendung einer Vielzahl von Termen resultierende hohe Dimensionalität des Merkmalsraumes für viele der zur Erstellung eines Klassifikators verwendeten Lernverfahren problematisch ist, trachtet man – wiederum mit verschiedenen Ansätzen – danach, die Größe des Vektorraums zu reduzieren.

Die induktive *Erstellung* eines rangordnenden Klassifikators für eine bestimmte Klasse besteht in der Definition einer Funktion, die bei Anwendung auf jeden Vektor einen Wert ausgibt, der meist zwischen „0" und „1" variiert und ausdrückt, zu welchem Grad das Dokument zur betreffenden Klasse gehört. Am Ende des Prozesses werden pro Klasse die Dokumente bzw. pro Dokument die Werte für die Klassen absteigend nach diesen Werten ranggeordnet. Die Erstellung eines *„harten"* Klassifikators basiert entweder auf einer Funktion, die eine Ja-Nein-Entscheidung trifft (das Dokument gehört zur betreffenden Klasse oder nicht) oder aber auf der Verwendung eines rangordnenden Ansatzes zuzüglich der Definition eines *Schwellenwertes*, der dann entscheidet, ob das Dokument zur Klasse gehört oder nicht. Auf die Vielzahl der methodischen Ansätze kann hier nicht näher eingegangen werden; erwähnt seien lediglich Bezeichnungen wie probabilistische Klassifikatoren, Entscheidungsbäume, Regressionsmethoden, Rocchio-Algorithmus (lineare Batch-Methode), Online-Methoden (inkrementelle Klassifikatoren), künstliche neuronale Netze, Instanz-basierte Klassifikatoren (z.B. k-Nearest-Neighbors-Verfahren), Support-Vektor-Maschinen. In letzter Zeit ist auch der kombinierte Einsatz verschiedener Klassifikatoren populär geworden.

4.2 Die LCC-Studie von Larson

Die 1992 publizierte Untersuchung von Larson,[9] in der über Experimente zur automatischen Klassifizierung von MARC-Datensätzen nach der *Library of Congress Classification* (LCC) berichtet wird, gilt als bahnbrechend, da sich zuvor – und auch danach [!] – niemand in ähnlich systematischer Weise mit der automatischen Zuteilung von Notationen eines bedeutenden Klassifikationssystems zu *Büchern* (repräsentiert durch bibliographische Datensätze) beschäftigt hat. Den Experimenten lagen rund 30.000 MARC-Katalogisate aus einer bibliothekswissenschaftlichen Kollektion zugrunde, die auf Grund dieser Herkunft zu 92 % eine Notation aus der LCC-Hauptklasse „Z" (Bibliography, Library Science and Information Science) aufwiesen. Sie verteilten sich auf 5.765 verschiedene Klassen aus „Z". Die dafür erstellten Klassendefinitionen (Vektoren von Attributgewichten) basierten auf allen *Dokumenten*, die zur jeweiligen Klasse gehörten, wogegen die Terme aus den Tafeln bzw. Registern der LCC *nicht* verwendet wurden. Die Experimente wurden mit einer Testmenge von 283 neuen, bereits nach der LCC klassifizierten, aber noch nicht in der Datenbank enthaltenen Katalogisaten durchgeführt. Für das automatische Klassifizieren dieser Dokumente wurde das probabilistische IR-System *Chesire* eingesetzt (d.h. der Klassifizierungsprozess wurde als Information-Retrieval-Prozess definiert). Ziel war es, für die neuen, ebenfalls durch Vektoren von Termgewichten repräsentierten Dokumente die jeweils beste Klasse zu finden bzw. die Übereinstimmung dieses Ergebnisses mit der „wahren", d.h. vorab intellektuell zugeteilten, Klasse zu prüfen.

Zur Bestimmung der Ähnlichkeitswerte für die Rangreihung der Klassen nach dem Grad ihrer Übereinstimmung mit einem Eingabedokument wurde das Skalarprodukt von Klassenvektor und Dokumentenvektor errechnet. Dabei wurden vier verschiedene Methoden der Berechnung von *Termgewichten* getestet:

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

(2) *TFIDF-Gewichtung*:[10] Diese bekannte Methode gibt den Termen, die in einem Dokument oft, in der gesamten Kollektion jedoch selten auftreten, die höchsten Gewichte, und denjenigen Termen, die im Dokument selten, in der Kollektion jedoch häufig auftreten, die niedrigsten Gewichte.

(3) *„Model 1C":* Dieses probabilistische Gewichtungsverfahren basiert auf der bedingten Wahrscheinlichkeit, mit der ein Klassifizierer für ein Dokument, das den Term T enthält, die Klasse K zuteilen würde. Dies wird auf

[9] *Larson*, R. R. (1992). Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science*. 43(2). S. 130–148.

[10] TFIDF = Termfrequenz * Inverse Dokumentenfrequenz.

der Basis der vorab klassifizierten Dokumente für jeden Term berechnet. Der Dokumentenvektor besteht aus binären Gewichten wie bei (1).

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

(5) Des Weiteren wurden fünf Varianten der *Attributauswahl* getestet:

a) All elements: Die Terme aus dem Titel und aus allen Library of Congress Subject Headings (LCSH);

b) Title and first subject: Titel und erster LCSH (die Regeln der Library of Congress besagen, dass die LCC-Notation auf der Grundlage des ersten LCSH vergeben werden soll);

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

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

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

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

i *Full stemming:* Die Stichwörter aus dem Titel- bzw. LCSH-Feld wurden unter Einsatz des Systems *SMART*[11] auf ihre Stammformen reduziert;

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

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

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

Aus der detaillierten Ergebnisdarstellung bei Larson seien hier nur die Hauptresultate angeführt:

- Die besten Ergebnisse – definiert als größter Anteil der „wahren" Klasse auf Rangplatz 1 – erzielte die Kombination von *weighted relative frequency matching* (4), *first subject heading* (d) und *plural stemming* (ii). Damit konnten 46,6 % der neuen Katalogisate korrekt zugeordnet werden. Eine Inspektion der Klassen „zweiter Wahl" ergab, dass viele davon eine akzeptable Alternative zu der durch die menschlichen Klassifizierer gewählten Klassen

[11] Vgl. *Salton*; G.; *McGill*, M. J. (1987). *Information Retrieval: Grundlegendes für Informationswissenschaftler*. Hamburg usw.: McGraw-Hill. [hier: S. 139].

waren. 74,4 % der „wahren" Klassen befanden sich unter den jeweils 10 bestgereihten.

- In ca. 45 % der mit diesem Verfahren korrekt klassifizierten Fälle enthielten auch die LCSH-Normdatensätze die exakte Spezifizierung der betreffenden Notation, während in den übrigen Fällen nur unvollständige Notationen, Bereiche von Notationen oder gar keine Notationen verzeichnet waren.
- Der Versuch, die LCSH in Form von *Phrasen* zu verwenden (iii), erbrachte keine guten Ergebnisse, da dieser Ansatz zwar viele nicht korrekte, aber auch viele korrekte Klassen zurückwies.
- Die auf den *TFIDF-Gewichten* basierende Matching-Methode (2) erbrachte auffallend *schlechte* Resultate. Larson führte diesen Befund auf die Begrenztheit des Vokabulars zurück, zumal die verwendeten MARC-Datensätze nur Titel und LCSH enthielten.
- Die verschiedenen Verfahren erbrachten sehr unterschiedliche Resultate in Bezug auf die jeweils korrekt klassifizierten Fälle. Eine Analyse dieser „best ranks" ergab, dass 76,3 % korrekte Zuordnungen erreichbar gewesen wären, wenn für jedes Buch das jeweils dafür am besten geeignete Verfahren wählbar gewesen wäre.
- Larson schloss daraus, dass eine vollautomatische Vergabe von LCC-Notationen nicht realistisch sei, aber eine semi-automatische Vorgangsweise, bei der – möglicherweise auch auf der Basis der Kombination einiger Verfahren – ein menschlicher Klassifizierer Vorschläge zur Auswahl dargeboten bekäme, weiterverfolgt werden sollte.

4.3 Weitere Untersuchungen

Das ACS-Verfahren von Cheng & Wu.[12] Die an der Hongkong Polytechnic University beheimateten Autoren berichten von einer Untersuchung mit der auch in Südostasien weit verbreiteten DDC. Das von ihnen erstellte *Automatic Classification System* (ACS) basiert auf dem Vektorraummodell und einem neu entwickelten Ähnlichkeitskoeffizienten. Für die Erstellung der Klassenrepräsentationen wurde nicht das Vokabular der DDC herangezogen, sondern jenes der Titel und Kapitelüberschriften einer Kollektion zuvor intellektuell klassifizierter Bücher. Dabei erfolgte die Entfernung von Stoppwörtern und Wörtern mit geringer Auftretensfrequenz sowie die Bereinigung von Synonymformen (Akronyme, Schreibvarianten, Singular-/Pluralformen), jedoch keine Stammformenbildung.

Mit dem erwähnten Ähnlichkeitsmaß wurde die Übereinstimmung eines zu klassifizierenden Buches mit dem Eigenschaftsvektor der jeweiligen Klasse geprüft. Bei über- bzw. untergeordneten Klassen wurde paarweise unter-

[12] *Cheng*, P. T. K.; *Wu*, A. K. W. (1995). ACS: An automatic classification system. *Journal of Information Science*. 21(4). S. 289–299.

sucht, mit welcher Hierarchieebene eine stärkere Assoziation vorlag; war dies die übergeordnete Klasse, so brach das Verfahren ab, war es die untergeordnete, so erfolgte eine weitere Iteration mit der nächsten Hierarchiestufe.

Bei den untersuchten Dewey-Klassen handelte es sich um „510" (Mathematics) und, als untergeordnete Klasse, „515" (Calculus and analysis). Für beide Klassen wurden aus den Beständen einer Hochschulbibliothek in Hongkong bereits nach der DDC klassifizierte Werke ausgewählt und in zwei Gruppen geteilt. Anhand der Bücher der ersten, größeren Gruppe wurden die Klassenrepräsentationen erstellt, während es sich bei den Büchern der zweiten Gruppe um Dokumente handelte, deren Vokabular nicht in diese Klassendefinitionen eingegangen war und die von *ACS* neu zu klassifizieren waren.

Die Resultate fielen für beide Gruppen mit 86,7 % bzw. 90 % korrekten Zuordnungen sehr vorteilhaft aus. Um das beste Ergebnis zu erzielen, war auch mit einer Heuristik experimentiert worden, die ein zu klassifizierendes Buch dann der „*parent class*" zuordnete, wenn der Assoziationskoeffizient für diese Klasse, multipliziert mit einem Schwellenwert, größer als der entsprechende Koeffizient für die untergeordnete Klasse war. Dabei wurden mit dem besten Schwellenwert (1,20) durchschnittlich 88,4 % korrekte Zuordnungen erreicht. Die Fehlklassifikationen wurden auf das Fehlen von Kapitelüberschriften, nicht aussagekräftige Titel, in generelleren Werken versteckte Spezialthemen und Fehler der intellektuellen Klassifizierung zurückgeführt. Für die praktische Anwendung wurde die Kombination mit menschlicher Intervention – Auswahl der Hauptklasse, bei der *ACS* beginnen sollte, Eingreifen bei problematischen Entscheidungen zwischen Klassen – empfohlen.

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

AutoBC. „Automatic Book Classification", ein von Kim & Lee (2002)[13] in Südkorea entwickeltes Verfahren, verwendet die auf S. R. Ranganathan zurückgehende *Colon Classification* (CC). Realisiert wurde vorerst nur ein Klassifikator für das Fachgebiet „Bibliothekswissenschaft". Grundmodul des Systems ist eine nach Fachgebieten und Facetten strukturierte Vokabulardatenbank; für das genannte Fachgebiet umfasst diese 387 aus der CC und der DDC extrahierte Begriffe. Kim & Lee hängen der eher fragwürdigen These an, wonach die Titel von Monographien alleine genügend Aussagekraft für eine automatische Klassifizierung besitzen; nur in Einzelfällen müssten von bibliothekarischer Seite zusätzlich Schlagwörter hinzugefügt werden. *AutoBC* ordnet das (Titel-)Vokabular der zu klassifizierenden Bücher auf Grund der in der Vokabulardatenbank vorgefundenen Strukturen und der Termfrequenzen den Fachgebieten, Facetten und Isolaten der CC zu und kombiniert diese

[13] *Kim*, J.-H.; *Lee*, K.-H. (2002). Designing a knowledge base for automatic book classification. *Electronic Library*. 20(6). S. 488–495.

nach der Facettenformel der CC zu gültigen Notationen. Ein Test mit 365 Büchern ergab, dass 81 % davon „klassifizierbar" waren und die restlichen nach einer weiteren Anreicherung der Vokabulardatenbank klassifiziert werden konnten. Details über die Güte dieser Zuordnungen sind ebenso wenig bekannt wie Einzelheiten über die tatsächliche Vorgangsweise im Rahmen dieses nur unzureichend dokumentierten Ansatzes.

ACN und UDC-AUTCS. Diese beiden in Japan erstellten Verfahren sollen Klassifizierer in Bibliotheken bei der Bildung von Notationen maschinell unterstützen. Das bereits in den 1980er Jahren entstandene Modul „Automatic Classification Numbering"[14] verwendete die *Nippon Decimal Classification* (NDC) und beruhte auf der interaktiven Eingabe von Sachbegriffen durch den Katalogisierer. Das System suchte diese Begriffe in einer Datenbank, die das Vokabular der NDC (Tafeln, Hilfstafeln und Register) beinhaltete und schlug Kandidaten-Notationen vor. Nach einer menschlichen Auswahlentscheidung bildete es unter Beachtung der NDC-spezifischen Verknüpfungsregeln die endgültige Notation in formal korrekter Form. Ein „Test" mit 24 [!] bereits manuell klassifizierten Büchern erbrachte eine „hohe" Übereinstimmungsrate. Das später für die UDK entwickelte „UDC Number Automatic Combination System"[15] arbeitete nach demselben Schema. Offensichtlich vermochte diese neuere Variante auf Grund von Eingabeparametern zu erkennen, ob der betreffende Begriff in den Haupt- oder in den Hilfstafeln nachgeschlagen werden sollte. Die endgültige Notation wurde nach einer neuerlichen Bearbeiterentscheidung auf Grund der Verknüpfungsregeln der UDK erstellt. Beide Ansätze haben jedoch wenig mit automatischer Klassifizierung im engeren Sinn zu tun und können im günstigsten Fall als Katalogisierungshilfe bewertet werden.

NDC und Bücher. Ishida[16] berichtete über ein Experiment, im Rahmen dessen japanische Bücher auf der Basis von Katalogisierungsdatensätzen automatisch den Klassen der NDC zugeordnet wurden. Dabei wurden verschiedene Extraktions- und Gewichtungsmethoden getestet. Der Studie lag eine Kollektion von 1.000 Büchern zugrunde; das beste erzielte Resultat lag bei 55,9 % korrekten Zuordnungen.[17]

Automatic Dewey Decimal Classification. Unter dieser Projektbezeichnung versucht eines der bekanntesten Fachinstitute Indiens (Documentation

[14] *Ishikawa*, T. (1988). The man-machine interface aspect of an automatic classification numbering system. *Journal of Information Processing*. 11(3). S. 199–205.

[15] *Ishikawa*, T.; *Nakamura*, H.; *Nakamura*, Y. (1994). UDC number automatic combination system (UDC-AUTCS): Implications for classifying and document[-like object] retrieval. In: *Knowledge Organization and Quality Management: Proceedings of the 3rd International ISKO Conference, Copenhagen, DK, 20–24 June 1994.* Hrsg.: *Albrechtsen*, H.; *Oernager*, S. Frankfurt/Main: Indeks Verl. (Advances in knowledge organization; 4). S. 328–333.

[16] *Ishida*, E. (1998). An experiment of automatic classification of books using Nippon decimal classification [Text in japanischer Sprache]. *Library and Information Science*. (39). S. 31–45.

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

Research & Training Centre, Bangalore), Bücher automatisch nach der DDC zu klassifizieren.[18] Das System basiert auf einem Parser für natürliche Sprache, einem Expertensystem und einem regelbasierten Algorithmus für die Notationsvergabe. Das der syntaktischen Analyse der Buchtitel durch den Parser zugrunde liegende Lexikon wurde auf Basis des „Relative Index" der DDC erstellt.[19]

5. Fazit: Automatisches Klassifizieren und Bibliothekskataloge

Das bislang offensichtliche Fehlen eines breiteren Interesses an der automatischen Klassifizierung von Buchbeständen ist überraschend. Mag es für den amerikanischen Raum noch dadurch erklärbar sein, dass neue MARC-Datensätze in der Regel bereits klassifiziert und dass auch bei retrospektiven Erfassungsprojekten auf Grund der großen Freihandaufstellungen häufig Aufstellungsnotationen verfügbar sind, so besteht im deutschen Sprachraum eine andere Situation. In den 1990er Jahren wurde errechnet, dass allein in Deutschland 52 Millionen älterer (Formal-)Katalogisate auf ihre Konvertierung warteten,[20] von denen inzwischen ein gewisser Teil maschinell erfasst, aber vermutlich *nicht* sachlich (klassifikatorisch) erschlossen sein dürfte.

Auf Grund der durchgeführten Recherchen scheint festzustehen, dass mit Ausnahme der inzwischen schon mehr als ein Jahrzehnt zurückliegenden Untersuchung von Larson keine *signifikanten* Studien oder Anwendungen aus dem Bibliotheksbereich vorliegen, die sich mit dem automatischen Klassifizieren von Büchern bzw. Katalogisaten beschäftigen. Enttäuschend ist in diesem Zusammenhang v.a. auch, dass sich nicht einmal das bibliothekarische Infrastrukturunternehmen OCLC dieses Themas angenommen hat. Das von OCLC durchgeführte Projekt zum automatischen Klassifizieren, *Scorpion*,[21] zielte ausschließlich auf Web-Dokumente ab; auch der im Rahmen des OCLC-Dienstes *CORC/Connexion* verfügbare Klassifikator dient nur der Katalogisierung elektronischer Ressourcen.[22]

So kann die von Gödert[23] geäußerte Hoffnung, wonach die wachsende Zahl der in OPACs verzeichneten, aber nicht inhaltlich erschlossenen Bücher

---

[18] *Srishaila*, S. (2001). Tools for assigning subjects to e-documents: A step towards organizing Internet resources. *Workshop on Multimedia and Internet Technologies, DRTC, Bangalore, 26–28 Feb.* WWW: <https://drtc.isibang.ac.in/handle/1849/83> [26. 06. 2005]. S. 7–8.

[19] Weitere Details zu diesem nur durch Sekundärliteratur bekannten Ansatz waren nicht zu ermitteln.

[20] *Beyersdorff*, G. (1993). Gesamtergebnisse und Empfehlungen (Kapitel 5). In: *Retrokonversion: Konversion von Zettelkatalogen in deutschen Hochschulbibliotheken: Methoden, Verfahren, Kosten*. Hrsg.: WEBER, K. Berlin: Deutsches Bibliotheksinstitut. S. 285–311.

[21] Zusammenfassende Darstellung: OBERHAUSER, op. cit. (2005), S. 79–98.

[22] Ibid.

[23] *Gödert*, W. (2002). „Die Welt ist groß – Wir bringen Ordnung in diese Welt". *Information – Wissenschaft und Praxis*. 53(7). S. 395–400. [hier: S. 396].

ein verstärktes Interesse an der automatischen Zuteilung von Notationen hervorrufen würde, auf der Basis der hier untersuchten Literatur vorerst nicht bestätigt werden.

Im Vergleich zur automatischen Klassifizierung von elektronischen Dokumenten, die ja zumeist im Volltext vorliegen, mutet die im Fall von Katalogisaten auf Wörter aus Titel, Untertitel und (allfällig) verbaler Sacherschließung begrenzte Textmenge sehr spärlich an. Dieser Umstand mag für Interessenten an automatischen Klassifizierungsverfahren abschreckend gewirkt haben. In diesem Zusammenhang sei auch an Larsons Befund erinnert, wonach das sonst so bewährte TFIDF-Gewichtungsverfahren bei *Katalogisaten* besonders schlechte Ergebnisse erbrachte. Zwar hat sich z. B. im Rahmen der *MILOS*-Projekte gezeigt, dass sogar auf einer solchen limitierten textuellen Basis mit einer linguistischen Methode des automatischen *Indexierens* eine sinnvolle Erweiterung des inhaltstragenden Wortschatzes erreicht werden kann,[24] doch steht der Beweis noch aus, dass dies auch für das automatische *Klassifizieren* gilt.

Somit muss festgehalten werden, dass – im Gegensatz zum automatischen Indexieren – gegenwärtig an einen praktischen Einsatz des automatischen Klassifizierens für Bibliothekskataloge noch nicht zu denken ist. Wie auch aus anderen Anwendungsbereichen bekannt ist, ist die durch die verschiedenen Verfahren erzielte Klassifizierungsgüte bis dato meist nicht zufrieden stellend. Viele Verfahren sind – auch wenn kommerzielle Vermarkter anderes versprechen mögen – dem Labor noch nicht entwachsen. Und vor allem: Die Zahl der bisher vorliegenden Untersuchungen ist viel zu klein, um brauchbare Schlüsse ziehen zu können.

Angesichts des Übersetzungsprojekts *DDC Deutsch*[25] besteht aber zumindest die Chance, dass im Falle der Akzeptanz der DDC als „Einheitsklassifikation" für deutschsprachige Online-Kataloge auch Bestrebungen zur Ausstattung eines möglichst großen Teils der darin enthaltenen Katalogisate mit Notationen aus diesem System entstehen könnten. Hier soll nicht weiter über die verschiedenen Möglichkeiten zur Übernahme solcher Notationen diskutiert werden; für eine automatische Klassifizierung stünden aber sicherlich Trainingsdokumente in großer Zahl zur Verfügung, wenngleich der Anteil deutschsprachiger Katalogisate unter den weltweit nach der DDC erschlossenen Titel eher gering sein dürfte.

Ein zweites Klassifikationssystem, für das große Zahlen bereits erschlossener Katalogisate – darunter auch sehr viele deutschsprachige – existieren, ist die *Basisklassifikation*,[26] die zudem in drei Sprachen (niederländisch, englisch, deutsch) vorliegt. Mit diesem System, das für bestimmte Funktionen in Bibliothekskatalogen durchaus sehr gut geeignet ist, zu experimentieren und

[24] *Sachse*, E.; *Liebig*, M.; *Gödert*, W. , op. cit. (1998).
[25] *Gödert*, W. , op. cit. (2002).
[26] http://www.gbv.de/du/sacher/bk3_gbv.shtml [25. 06. 2005]

dabei auch wichtige Erfahrungen in methodischer Hinsicht zu gewinnen, würde sich gewiss lohnen.

Jüngste Bestrebungen im deutschsprachigen Bereich inkludieren ein Projekt im Rahmen des Lehrbetriebes am Institut für Informationswissenschaft der Fachhochschule Köln.[27] Dabei sollen zwar nicht Katalogisate aus einem OPAC, aber immerhin solche aus einer deutschen sozialwissenschaftlichen Datenbank automatisch klassifiziert werden. Den dabei gemachten Erfahrungen, insbesondere was die Auswahl des Klassifikators sowie der eingesetzten linguistischen Verfahren zur Aufbereitung des deutschsprachigen Materials betrifft, darf jedenfalls mit Spannung entgegengesehen werden – wie natürlich auch der erzielten Klassifizierungsgüte.

[27] *Gödert*, W.: Persönliche Mitteilungen an den Verfasser, Mai–Juni 2005.

Bibliotheksbuchbinderei
Werner Schober

Am 29.08.1890 gründete mein Urgroßvater Josef Nagel, in 1050 Wien, Rüdigergasse 16 eine Buchbinderei sowie eine Papierhandlung mit Schreib- und Zeichenrequisiten, gewerblichen Drucksorten und einer Lizenz zum Verkauf von Schul- und Gebetsbüchern, Kalendern und Heiligenbildern.

Im Jahre 1895 wurde der Standort der Firma nach 1040 Wien, Paulanergasse 12 verlegt. Der Betrieb wurde von meinem Großvater mit bis zu 40 Mitarbeitern geführt. Im Jahre 1945 wurde bedingt durch die Kriegsereignisse ein staatlicher Kommissär eingesetzt, welcher gemeinsam mit meiner Mutter Ottilie Glaser die Firma als Verlagsbuchbinderei weiterführte.

Im Jahre 1965 übernahm ich im Alter von 21 Jahren als jüngster Buchbindermeister die Einzelfirma und mein Großvater ging als bereits 82jähriger in Pension. Da der Betrieb zu diesem Zeitpunkt keine Mitarbeiter mehr hatte, mußte ich mit dem Aufbau der Bibliotheksbuchbinderei von Neuem beginnen. Mit Hilfe von gutem Fachpersonal verlief dies sehr zufriedenstellend.

Seit dem Jahre 1993 leitet mein Sohn Mag. Ing. Johannes Schober die Bibliotheksbuchbinderei und ich blicke mit Freude auf stolze 115 Jahre Firmenbestehen im Familienbesitz zurück.

1040 Wien, Paulanergasse 12, Tel.: 581 46 32, Fax: 257 32 22 23, www.schogla.com