Perspektive Bibliothek 3.1 (2014), S. 85-110

Automatisierte und semiautomatisierte Klassifizierung: Eine Analyse aktueller Projekte

Anna Kasprzik

Bibliothek der Universität Konstanz / Bibliotheksakademie Bayern[1]

anna.kasprzik@googlemail.de

1. Einleitung und Begriffsklärungen

Die inhaltliche Erschließung von Medien an Bibliotheken dient primär dazu, dem Nutzer die thematische Orientierung und das Auffinden der gesuchten Informationsquellen im Bestand zu erleichtern. Diese Erschließung wurde bis vor einigen Jahrzehnten ausschließlich intellektuell vorgenommen, das heißt, mit Hilfe des Welt- und Fachwissens der bearbeitenden Personen. Die fortschreitende Verlagerung von textueller Information und von Metadaten zu diesen Texten ins digital Verfügbare und die sich verschärfende Zeit- und Geldknappheit legen den Einsatz von halb- oder vollautomatischen Verfahren der Inhaltserschließung zur Entlastung des Bibliothekspersonals, insbesondere der Fachreferenten, zugunsten anderer Aufgaben nahe. Auch besteht die Hoffnung, mittels solcher Methoden der mit der digitalen Verfügbarkeit einhergehenden explosionsartigen Zunahme von Informationsquellen Herr zu werden.

In der computerbasierten Inhaltserschließung unterscheidet man *halbautomatische* Verfahren, die lediglich eine Reihe von Vorschlägen zur Beschreibung eines Dokuments generieren, von *vollautomatischen* Verfahren, bei denen die Zuordnung direkt vorgenommen wird. Des Weiteren unterteilt man in *statistische*

[1] Dieser Artikel ist als eine Arbeit im Rahmen der Referendarsausbildung für die vierte Qualifikationsebene an der Bibliotheksakademie Bayern entstanden.

CC BY

und *linguistische* Verfahren.[2] Statistische Verfahren stützen sich auf die Auftretenshäufigkeit eines Begriffes, um über seine Eignung zur Charakterisierung eines Dokuments zu entscheiden. *Entscheidungsstark* sind Begriffe, die allgemein bzw. im vorliegenden Textkorpus nicht so häufig auftreten (dies schließt sogenannte „Stoppwörter", also etwa Artikel, Pronomen, Präpositionen oder Adverbien aus), im zu beschreibenden Dokument aber mit mittlerer bis hoher Häufigkeit.[3] Linguistische Verfahren berücksichtigen diverse Gesetzmäßigkeiten der Sprache, um geeignete Begriffe zur Beschreibung zu eruieren. So abstrahieren sie von Flexion, Derivation und Komposition, um verschiedene grammatikalische Auftretensformen auf ein und dasselbe Konzept zurückzuführen, erkennen ganze Wortgruppen als zusammengehörig oder setzen auf der semantischen Ebene Synonyme in Beziehung. Dies geschieht auf der Basis von allgemein formulierten linguistischen Regeln und/oder Wörterbucheinträgen, in denen alle Auftretensformen verzeichnet sind.[4] Die Unterscheidung in statistische und linguistische Verfahren schließt ein gemeinsames Auftreten in der Praxis nicht aus: Linguistische Verfahren bieten sich an, um einen Text als Input für ein statistisches Verfahren aufzubereiten, und generell werden oft mehrere Strategien zu einem Verfahren kombiniert.[5]

Interessanterweise stehen laut Gödert et al. (2012) intellektuelle und automatische Verfahren nicht in Konkurrenz zueinander, da erstere den Anspruch zu erfüllen suchen, ein Dokument möglichst zutreffend zu beschreiben, während letztere rein dem Zweck einer Verbesserung der Auffindbarkeit (also einer Erhöhung der Zahl der Zugriffsmöglichkeiten) dienen sollen.[6] Überspitzt formuliert hieße das, wenn eine automatisch vergebene Beschreibung das Dokument völlig unzutreffend charakterisiert, aber zu einer hohen Auffindbarkeit führt (denkbar etwa bei Sachverhalten, zu denen in der Bevölkerung ein weit verbreiteter Irrglaube existiert), ist sie im Sinne der automatischen Inhaltserschließung legitim. Trotzdem gilt im Allgemeinen natürlich auch für letztere das Desiderat der *Konsistenz* – gleiche Sachverhalte sollten stets gleich beschrieben werden, sodass eine völlige Willkür weitgehend ausgeschlossen wird.

Das Finden und die Zuordnung geeigneter Begriffe zur kompakten inhaltlichen Repräsentation eines Textdokuments nennt man *Indexierung*.[7] Aus dem Ergebnis dieses Vorgangs lässt sich ein *Index* erstellen, der die Fundstellen der als

[2] Vgl. z.B. Gödert et al. (2012, S. 246).
[3] Vgl. Gödert et al. (2012, S. 257–260, 291–293).
[4] Vgl. z.B. Gödert et al. (2012, S. 260–263, 285–290).
[5] Vgl. Gödert et al. (2012, S. 246).
[6] Vgl. Gödert et al. (2012, S. 246).
[7] Vgl. z.B. Siegmüller (2007, S. 11).

relevant eingestuften Begriffe verzeichnet.[8] Falls mit statistischen Methoden gearbeitet wurde, kann im Index auch der Relevanz*grad* eines Begriffes in einem bestimmten Dokument vermerkt sein[9] und bei halbautomatischen Verfahren wiederum einem menschlichen Bearbeiter als Entscheidungsgrundlage dienen. Man unterscheidet *kontrollierte* Indexierung, bei der die ausgewählten Begriffe einem festen Thesaurus, Schlagwortkatalog oder einer Klassifikation entstammen, von *freier* Indexierung mit nicht vorgegebenen Begriffen.[10] Die Begriffe können dem vorliegenden Material aus dem Textdokument entstammen (*Stichwörter*), müssen es aber nicht (allgemein: *Schlagwörter*). Ein Beispiel: Ein Stichwort in Bezug auf die Phrase „schwarzes Gold" wäre „Gold", ein (inhaltlich treffenderes) Schlagwort wäre hingegen „Erdöl". Eine Suche nach Begriffen, die *nicht* aus dem Material hervorgehen, muss jedoch gezwungenermaßen mit aufwändigen semantischen Methoden erfolgen und stellt für die automatische Indexierung eine Herausforderung dar.[11]

Die Qualität der Sacherschließung mit maschinellen Verfahren verbessert sich erwartbarerweise, je mehr Information über ein Textdokument in elektronischer Form vorliegt. Im Idealfall ist das der durchsuchbare Volltext, Abstufungen davon sind (repräsentative) Textauszüge, Abstracts und Inhaltsverzeichnisse bis hin zu herkömmlichen bibliothekarischen Metadatensätzen.

In dieser Arbeit konzentrieren wir uns auf die *klassifikatorische* Sacherschließung, also die Einordnung eines Textdokuments in eine gegebene *Systematik* (*Klassifikation*) bzw. die Vergabe von *Notationen* als Repräsentanten von Klassen aus dieser Systematik.[12] Der allergrößte Teil der aktuellen Projekte zur (fachübergreifenden) automatischen Klassifizierung bezieht sich auf die international weit verbreitete *Dewey-Dezimalklassifikation* (DDC), die seit ihrer Übersetzung (2005) auch im deutschsprachigen Raum zunehmend zur Anwendung kommt.[13] Ein weiteres prominentes Projekt (Abschnitt 2.1) befasst sich mit der *Regensburger Verbundklassifikation* (RVK), welche an zahlreichen wissenschaftlichen Bibliotheken zur Aufstellung von Freihandbeständen genutzt wird und darüber hinaus in der kooperativen Sacherschließung Verwendung findet.[14]

[8] Vgl. z.B. Gödert et al. (2012, S. 248).
[9] Vgl. Siegmüller (2007, S. 28).
[10] Vgl. z.B. *Wikipedia – Indexierung* (o.D.).
[11] Ein Ansatz zu diesem Problem sind sogenannte *Expertensysteme*, vgl. Oberhauser (2005, S. 22).
[12] Siehe Gödert et al. (2012, S. 69, 355–356): „Eine *Systematik*, auch *Klassifikation* genannt, strukturiert Inhalte in hierarchisch angeordnete *Klassen*. Hierbei fasst eine Klasse Begriffe oder Konzepte zusammen, die über mindestens ein gemeinsames Merkmal verfügen" (Hervorhebungen durch die Autorin).
[13] Siehe auch *Dewey-Dezimalklassifikation* (o.D.).
[14] Siehe auch *Regensburger Verbundklassifikation* (o.D.).

Durchgesetzt haben sich im Bereich der automatischen Klassifizierung Methoden aus dem Bereich des *maschinellen Lernens*, einer Unterdisziplin der *künstlichen Intelligenz*. Solche Verfahren bestehen in der Regel aus einer Phase des *Trainings* (oder Erfahrungssammlung) anhand von intellektuell erstellten bzw. geprüften Präzedenzfällen und einer *Klassifizierungsphase*, in der dann noch nicht klassifizierte Dokumente hinsichtlich ihrer Charakteristika analysiert und in die Zielklassifikation eingeordnet werden (bzw. eine nach Relevanz gerankte Liste von Klassen für sie ausgegeben wird).[15] Als mögliche Klassifizierungsfunktionen seien hier insbesondere die folgenden zwei Typen genannt: *Instanzbasierte* (oder *fallbasierte*) *Klassifikatoren*, welche sich an all denjenigen vorliegenden Präzedenzfällen orientieren, die dem zu klassifizierenden Dokument am ähnlichsten sind (bei einer exakt zu bestimmenden Definition von „ähnlich"), und sogenannte *Support-Vektor-Maschinen* (SVMs) – wenn Dokumente als Vektoren repräsentiert sind, in denen zu jedem vorkommenden Begriff seine Relevanz für die Bedeutung des Dokuments eingetragen ist, so berechnet eine SVM eine Hyperebene, welche den von diesen Vektoren aufgespannten Raum durch eine größtmögliche Lücke in zwei Klassen trennt. Die dieser Ebene am nächsten liegenden Datenpunkte nennt man *Support-Vektoren*, nur sie sind letztendlich relevant für die Unterscheidung der beiden Klassen. Ein neues Dokument wird dann in die eine oder in die andere Klasse eingeordnet abhängig davon, auf welcher Seite der Hyperebene es sich befindet.[16]

Zur Bewertung eines Klassifikators gibt es verschiedene Maße, am weitesten verbreitet sind die *Genauigkeit* (oder *Zuverlässigkeit*, engl. *precision*), welche angibt, mit welcher Wahrscheinlichkeit das einer Klasse zugeordnete Beispiel tatsächlich zu dieser Klasse gehören sollte, und die *Vollständigkeit* (oder *Komplettheit*, engl. *recall*), welche angibt, mit welcher Wahrscheinlichkeit ein Beispiel, das einer Klasse zugehören sollte, ihr auch zugeordnet wird. Häufig wird noch das sogenannte *F-Maß* angegeben, welches sich aus dem gewichteten harmonischen Mittel von Genauigkeit und Vollständigkeit ergibt.[17]

Schon in den 90er-Jahren gab es erste namhafte Projekte zur automatischen Klassifizierung, als Beispiele seien genannt: DESIRE (EU-Projekt, Schweden), GERHARD (DFG, Deutschland) und Scorpion (OCLC, USA).[18] In dieser Arbeit konzentrieren wir uns auf aktuelle Projekte (2007–2012) aus dem deutschsprachigen Raum. Wir greifen für jedes Projekt insbesondere die folgenden Punk-

[15] Vgl. Oberhauser (2005, S. 18, 20).
[16] Vgl. z.B. Oberhauser (2005, S. 30).
[17] Vgl. z.B. Oberhauser (2005, S. 32, 34) und *Wikipedia – Beurteilung eines Klassifikators* (o.D.).
[18] Vgl. z.B. Oberhauser (2005, S. 41, 67 und 79).

te heraus: Einige Eckdaten, die gewählte Methode, die gewählte Datengrundlage, Tests und Ergebnisse und eventuell im Zuge der Durchführung identifizierte Problembereiche. Eine zusammenfassende Analyse der inhaltlichen und organisatorischen Hauptaspekte findet sich in Abschnitt 3.

2. Beschreibung einiger aktueller Projekte

2.1 UB Mannheim: Das Projekt von Magnus Pfeffer (RVK)

Die Durchführung dieses Projektes geschah hauptsächlich im Rahmen der Masterarbeit von Magnus Pfeffer, die im Jahr 2007 vorgelegt wurde.[19] Von den 1,2 Millionen Titeln der UB Mannheim waren Anfang 2007 etwa die Hälfte durch Fremddatenübernahme mit einer oder mehreren Notationen aus der RVK versehen. Von einer vollständigen Annotation aller Titel im Katalog mittels eines automatisierten Verfahrens erwartete man sich die Unterstützung der Fachreferenten bei der Retrosystematisierung, eine Datengrundlage für die Abschätzung des Platzbedarfs der einzelnen Systemstellen und die Möglichkeit einer Online-Darstellung in einem „virtuellen Bücherregal".[20]

Die maschinelle Lernmethode der Wahl war das in der Einleitung erwähnte *fallbasierte Schließen*, bei dem zu jedem neuen Fall der ähnlichste aus einer gegebenen Fallbasis gesucht und die dort verzeichnete Lösung auch auf den neuen Fall angewandt wird. Auf das Szenario der UB Mannheim übertragen handelt es sich bei den Fällen um Titelaufnahmen und bei den Lösungen dafür um die vergebenen RVK-Notationen. Essentiell für die Methode ist natürlich eine exakte Definition von „Ähnlichkeit", welche sich in verschiedenen mathematischen Funktionen manifestieren kann.[21]

Als Grundlage dienten MAB2-Verbundabzüge von Titeldaten, aus denen die Titel- und Schlagwörter extrahiert wurden.[22] Die RVK-Klassen wurden aus der von der UB Regensburg bereitgestellten XML-Darstellung gewonnen und Klassen mit nicht-inhaltlichen Definitionen ausgeschlossen. Sodann wurden die Daten linguistisch aufbereitet (Entfernung von Stoppwörtern, Zerlegung in Lexeme)[23] und

[19] Pfeffer (2007a).
[20] Vgl. Pfeffer (2007b, S. 21).
[21] Vgl. Pfeffer (2007a, S. 22–25).
[22] Anfangs handelte es sich hier nur um Titeldaten aus der UB Mannheim, dies wurde später zu Gesamtabzügen der Verbünde SWB und HeBIS erweitert, siehe Pfeffer (o.D., Juni 2008, April 2009).
[23] Titel in verschiedenen Sprachen (Deutsch, Englisch) wurden getrennt behandelt. Deutsche Titelwörter wurden mit Hilfe des Tools *Morphy* zerlegt, nicht-deutsche mit *Snowball*, siehe Pfeffer (2007a, S. 17).

es wurden separate Indizes für die Titelwörter, die Identnummern der Schlagwörter und die Lexeme erstellt.[24]

Zur Klassifizierung eines neuen Titels wurden mittels der Indizes Titel mit übereinstimmenden Lexemen oder Schlagwörtern in der Fallbasis gesucht und dann anhand der Ähnlichkeitsfunktion verglichen. Das Verfahren liefert aus der Fallbasis die Menge der Titel mit einer Ähnlichkeit größer Null zum Kandidaten zurück, und für diesen können dann je nach Zielsetzung die Notationen der n ähnlichsten Titel vergeben werden.[25]

In den folgenden Testläufen wurden Titel aus der Fallbasis entfernt und mit Hilfe des entwickelten automatischen Verfahrens erneut klassifiziert. Hierbei erzielte die Berücksichtigung von sowohl Schlagwörtern als auch Lexemen in Verbindung mit der sogenannten *Hamming-Ähnlichkeit*[26] und der Übernahme aller Notationen des ähnlichsten Titels die besten Ergebnisse.[27] Als Vergleichswert wurde der Abstand von intellektuell und automatisch vergebenen Notationen in der Baumdarstellung der RVK genommen, wobei eine Übereinstimmung als „perfekt" und ein Abstand von 1–3 Knoten noch als „gut" bewertet wurde. Die oben genannte Kombination klassifizierte circa die Hälfte der Fälle „perfekt" und circa ein Viertel noch „gut".[28]

Der auf diese Weise neu und einheitlich klassifizierte Gesamtbestand der UB Mannheim wurde als systematischer Zugang zu den Medien in den OPAC eingespielt und für diverse Arbeiten im Fachreferat genutzt.[29] Das Verfahren wurde in den folgenden Jahren von Magnus Pfeffer weiterentwickelt und verbessert.[30,31]

---

[24] Vgl. Pfeffer (2007b, S. 28).
[25] Vgl. Pfeffer (2007b, S. 29).
[26] Die Hamming-Ähnlichkeit s_h („*similarity*") bezieht sich auf die Anzahl der sich unterscheidenden Wörter in zwei zu vergleichenden Texten A und B, in einer Formel ausgedrückt hieße das

$$s_h = 1 - \frac{|M(A) \cup M(B)| - |M(A) \cap M(B)|}{N(A) + N(B)}.$$

Hierbei seien $N(A)$ und $N(B)$ die Längen dieser Texte (also die Zahl der Wörter) und $M(A)$ und $M(B)$ die Mengen der vorkommenden Wörter (also bereinigt von Mehrfachvorkommen).
[27] Interessanterweise brachte eine Zerlegung der Titelwörter in Lexeme nur wenige Prozentpunkte Verbesserung gegenüber der Version mit unzerlegten Titelwörtern, vgl. Pfeffer (2007b, S. 33).
[28] Vgl. Pfeffer (2007b, S. 33).
[29] Vgl. Pfeffer (2007a, S. 28–29); Pfeffer (2007b, S. 34).
[30] Siehe auch Pfeffer (o.D.).
[31] Aufbauend auf diesem Projekt entstand auch eine BA-Arbeit zu maschinell generierten Korrelationen zwischen der RVK und der *Schlagwortnormdatei* (SWD), siehe Probstmeyer (2009) und Abschnitt 2.7.

Zu der hier beschriebenen Methode ist noch zu sagen, dass sie auf eine semantische Analyse und die Bildung allgemeiner Regeln verzichtet.[32] Damit ist sie angewiesen auf die Qualität der Daten in der verwendeten Fallbasis. So müssen die Notationen aus dem Verbundabzug korrekt und vollständig sein, rein formale Notationen („Zeitschrift") müssen ausgeschlossen werden, und einem Titel sollten nicht mehrere inhaltlich weit auseinanderliegende Notationen zugeordnet sein.[33] Ein weiteres Problem ist die Tatsache, dass die Komplexität einer Suche mit der Größe der Fallbasis signifikant steigt.[34] Pfeffer zieht in seiner Master-Arbeit das Fazit, dass der Anteil der schlecht oder falsch klassifizierten Titel für eine ungeprüfte Übernahme noch zu hoch ist[35] und zumindest durch den Einsatz heuristischer Methoden verbessert werden muss.[36,37]

2.2 DNB: Das Projekt PETRUS (DDC)

Für das Projekt „Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek" (PETRUS; Laufzeit 2009–2011)[38] wurden zu Beginn vier Anwendungsszenarien definiert, die alle zukünftig mit automatischen Methoden realisiert werden sollten: Die Erkennung paralleler oder ähnlicher Ausgaben und der Austausch von Metadaten zwischen diesen Titeldatensätzen; die Generierung von Datensätzen in der *Personennormdatei* (PND) beim Import neuer Titel und die Verknüpfung von Personennamen und Titeldatensätzen; die Einordnung in die DDC[39] und die Vergabe von Schlagwörtern auf der Grundlage der *Schlagwortnormdatei* (SWD).[40] Man konzentrierte sich zunächst auf die seit 2010 in der *Reihe O* getrennt erfassten Online-Publikationen, hatte aber bei der Entwicklung perspektivisch auch andere Publikationsformen im Blick.[41] Wir beschränken uns auf die Sachgruppenvergabe aus der DDC, welche an der DNB seit 2012 für deutsche und englische Netzpublikationen komplett maschinell erfolgt.[42]

[32] Vgl. Pfeffer (2007a, S. 11).
[33] Vgl. Pfeffer (2007b, S. 24).
[34] Vgl. Pfeffer (2008, Folie 6).
[35] Hochspezielle Titel wurden oft völlig falsch eingeordnet. Auf den für die Fachreferenten generierten Listen konnten diese jedoch intellektuell schnell identifiziert werden und es landeten nur circa 15% der Titel im falschen Fach, siehe Pfeffer (2007b, S. 28, 33–34). Für die halbautomatische Klassifizierung eignet sich das Verfahren also gut.
[36] Vgl. Pfeffer (2007a, S. 30).
[37] Siehe auch Pfeffer (o.D., Juni 2009).
[38] Siehe *PETRUS - Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek* (o.D.).
[39] Die von der DNB vergebenen Sachgruppen basieren auf den hundert Sachgruppen der zweiten Ebene der DDC, siehe Mödden & Tomanek (2012, S. 17).
[40] Vgl. Schöning-Walter (2011, S. 32).
[41] Vgl. Schöning-Walter (2011, S. 31).
[42] Vgl. Mödden & Tomanek (2012, S. 17).

Für Untersuchungen zur Machbarkeit wurden vier kommerzielle Softwareprodukte getestet, von denen letztendlich die *Averbis Extraction Platform* der Firma Averbis überzeugte.[43] Dieses System stellt verschiedene Tools zur linguistischen Vorverarbeitung und anschließenden Klassifizierung bereit. Als Datengrundlage dienen Volltexte, digitalisierte Inhaltsverzeichnisse von Printpublikationen und bibliographische Metadatensätze. Bei der Vorverarbeitung werden zunächst die Dokumentsprache, Satz- und Wortgrenzen, Wortarten, Nominalphrasen und Stoppwörter erkannt, danach erfolgt eine morphologische Zerlegung und eine semantische Analyse mit Bezug auf die SWD. Der Text wird als ein in mehreren Schritten aufbereiteter Vektor von nach Relevanz gewichteten Begriffen oder Lexemen dargestellt und dem Klassifikator übergeben.[44]

Die Sachgruppenvergabe mit Hilfe des Averbis-Klassifikators erfolgt mit dem in der Einleitung erwähnten Modell der *Support-Vektor-Maschinen* (SVMs), welche sich für große Dokumentenmengen und zahlreiche mögliche Vektoreinträge als besonders geeignet erwiesen haben. Als Trainingsdaten dienen intellektuell erschlossene Publikationen, um die Trennebene zwischen den Klassen daraus abzuleiten. Ein neu zu klassifizierendes Dokument wird dann je nach seiner Position in Bezug auf diese Ebene in eine Klasse eingeordnet und der Abstand zur Ebene als Konfidenzwert[45] vermerkt.[46]

In der Testphase standen etwa 45 000 digitale Volltexte zur Verfügung, überwiegend Hochschulschriften. Problematisch war die ungleiche Verteilung der Testdokumente auf die Sachgruppen – letztendlich konnten 81 der 101 Sachgruppen berücksichtigt werden, für die jeweils mindestens 70 Beispielobjekte vorhanden waren. Mit diesen Daten wurden dann jedoch bis zu 80% der Dokumente automatisch richtig klassifiziert.[47]

Mit Ablauf der Projektlaufzeit wurde das Verfahren als ins System eingebetteter Webservice in den laufenden Betrieb übernommen. Der Webservice holt sich den Text und die zugehörigen Metadaten und stellt anhand gewisser Schwellenwerte fest, ob sich das Dokument für die automatische Klassifizierung eignet. Sodann wird der Text mit der Averbis Extraction Platform klassifiziert und die resultierende Notation samt dem Konfidenzwert in den zugehörigen Datensatz ein-

[43] Siehe *Averbis Textanalyse* (o.D.).
[44] In einem ersten Schritt werden Einträge aus dem Vektor entfernt, die nur einen geringen Beitrag zur Klassifizierung beitragen, in einem weiteren Schritt werden die Einträge nach Häufigkeit gewichtet und Ausreißer neutralisiert, siehe Mödden & Tomanek (2012, S. 18–19, 21).
[45] Je größer der Abstand, desto sicherer die Einordnung.
[46] Vgl. Mödden & Tomanek (2012, S. 19–20).
[47] Vgl. Mödden & Tomanek (2012, S. 18).

getragen. Ist der Konfidenzwert zu niedrig, so wird dies ebenfalls vermerkt und eine Nachbearbeitung des Datensatzes ermöglicht.[48] In jedem Datensatz wird die Herkunft der enthaltenen Metadaten hinterlegt, sodass nachvollziehbar ist, ob die formalen und inhaltlichen Beschreibungen in der Nationalbibliographie intellektuell oder automatisch erstellt oder aus anderen Quellen übernommen wurden.[49] Durch die Einbeziehung dieser Angaben und die Anreicherung des Trainingskorpus mit neuen Beispielen soll der Klassifikator laufend optimiert werden.[50] Mit der steigenden Leistungsfähigkeit des Systems könnten Datensätze immer wieder bearbeitet werden, sodass eine intellektuelle Nachbesserung überflüssig wird.[51]

Das Qualitätsziel war eine Übereinstimmung von intellektuell und automatisch vergebenen Notationen von mindestens 80%.[52] Dieses Ziel wurde knapp erreicht, allerdings gab es große Unterschiede zwischen den einzelnen Sachgruppen, da manche Themengebiete nur schwer mit rein mathematischen Methoden voneinander abzugrenzen sind und verschiedene Fächer auch recht unterschiedliche Textmuster aufweisen.[53]

2.3 TIB Hannover: Das Projekt LINSearch / LINSearch 2

Das Projekt *LINSearch* (2007–2009) hatte zum Ziel die Entwicklung eines Systems zur maschinellen Indexierung deutsch- und englischsprachiger Texte aus Naturwissenschaften und Technik.[54] Die indexierten Daten sollten automatisch einem der sechs Fächer der TIB Hannover oder der Kategorie „Weitere Fächer" zugeordnet werden, um innerhalb des Rechercheportals *GetInfo* eine facettierte Suche nach Fach zu ermöglichen.[55] Die zunächst im Rahmen des Projekts entwickelte Methode, welche linguistische und statistische Verfahren miteinbezog, erreichte eine Genauigkeit von 70%,[56] spätere Testanwendungen mit einer kommerziellen

[48] Vgl. Mödden & Tomanek (2012, S. 20–21).
[49] Vgl. Schöning-Walter (2011, S. 31).
[50] Vgl. Mödden & Tomanek (2012, S. 23–24).
[51] Vgl. Schöning-Walter (2011, S. 31).
[52] Vgl. Schöning-Walter (2011, S. 34). Mit Beginn des Produktionsbetriebes sollte ein F-Maß von mindestens 70% über alle Sachgruppen und mittelfristig ein F-Maß von 70% für jede einzelne Sachgruppe erreicht werden, siehe Mödden & Tomanek (2012, S. 24).
[53] Vgl. Mödden & Tomanek (2012, S. 24).
[54] Beteiligt waren die TIB Hannover, das Forschungszentrum L3S, das FIZ Technik und das Institut der Gesellschaft zur Förderung der angewandten Informationsforschung der Universität des Saarlandes. Gefördert wurde das Projekt vom Bundesministerium für Wirtschaft und Technologie und die Projektlaufzeit betrug 2,5 Jahre, vgl. *LINSearch: Linguistisches Indexieren und Suchen* (o.D.).
[55] Soweit nicht anders angegeben, bezieht sich dieser Abschnitt auf die Quelle *Was lange währt… : Automatische Fächerklassifizierung in GetInfo über die Facette "Fach"* (2012).
[56] Vgl. Bähr (2010, Folie 13–18).

Plattform der Firma Recommind[57] und der auch von der DNB genutzten SVM-basierten Averbis Extraction Platform (siehe Abschnitt 2.2) erzielten 75% und bis zu 85%.[58] Letztendlich erwies sich allerdings ein mehrstufiges Verfahren zur Anreicherung der Metadatensätze als der verlässlichste Ansatz, bei dem nur die letzte Stufe einer automatischen Klassifizierung im eigentlichen Sinne entspricht: In Stufe 0 werden ganze Datenkollektionen pauschal einem TIB-Fach zugeordnet. In Stufe 1 werden schon vorhandene Notationen, unter anderem aus der *Basisklassifikation* (BK), der DDC, RVK und der lokalen Systematik, auf die sechs Fächer der TIB abgebildet.[59] In Stufe 2 werden Zeit- und Kongressschriften anhand der ZDB-Sachgruppen eingeordnet. In Stufe 3 schließlich wird mit Hilfe der Averbis Extraction Platform klassifiziert. Die Klassifizierung wurde an einer auf der BK basierenden Mappingtabelle trainiert und bezieht bestimmte Metadaten wie etwa die Titelfelder oder den Abstract mit ein. Eine Herausforderung war dabei der aus ihrer weit gestreuten Herkunft resultierende stark variierende Umfang und Erschließungsgrad dieser Daten.

Die Metadatensätze an der TIB werden über Massenroutinen importiert bzw. aktualisiert, in die die vier Klassifizierungsstufen nun direkt integriert sind. Je nach Eignung der Daten kommt die höchstmögliche Stufe zur Anwendung. Die verwendete Stufe, die resultierende Zuordnung zu einem oder zu mehreren Fächern und im Falle von Stufe 3 auch der Konfidenzwert werden anschließend in den Datensatz eingetragen.[60] Bei einem zu niedrigen Konfidenzwert, bei fehlender Sprachangabe oder nichtvorhandener Metainformation auf Deutsch oder Englisch und generell bei zu spärlichen Angaben wird der Datensatz der Kategorie „Weitere Fächer“ zugeordnet.

Als Verteilung auf die gestaffelten Stufen ergibt sich, dass tatsächlich nur ein Fünftel der Datensätze automatisch durch den Algorithmus in Stufe 3 klassifiziert wird, dafür aber zwei Drittel in Stufe 1, also durch die Nachnutzung klassischer intellektueller Erschließung. Stufe 3 kommt damit vor allem bei denjenigen Daten zum Zuge, die in der Regel gar nicht von Fachreferenten bearbeitet werden.[61]

[57] Siehe auch *Recommind CORE Platform* (o.D.). Diese Technologie basiert auf diversen patentgeschützten Verfahren wie z.B. dem Probabilistic-Latent-Semantic-Analysis-Algorithmus (PLSA).
[58] Vgl. Mensing (2011, Folie 10).
[59] Die TIB hatte zu der Zeit 40 Mio. Datensätze im Index, von denen etwa 20% mit Sacherschließungselementen versehen waren, siehe Mensing (2011, Folie 4).
[60] Es kommt nur eine Stufe zur Anwendung – eine denkbare Überprüfung der Stufen 0–2 durch Stufe 3 wurde aufgrund fehlender relevanter Ergebnisse verworfen.
[61] Vgl. Mensing (2011, Folie 15).

Laut einer Pressemitteilung[62] hat die Plattform bereits mindestens 1 Mio. Datensätze erfolgreich klassifiziert und wird seit 2012 produktiv eingesetzt. Die TIB plant den Ausbau aller vier Stufen für eine feinere Zuordnung innerhalb der einzelnen Fächer.

2.4 UB Bielefeld: Automatische Anreicherung von OAI-Metadaten (DDC)

Im Projekt „Automatische Anreicherung von OAI-Metadaten" (2009–2011)[63] sollten in Online-Repositorien frei zugängliche und von der *Open Archives Initiative* (OAI)[64] mit Metadaten erschlossene wissenschaftliche Publikationen einheitlich mit DDC-Notationen versehen werden, sodass die Datensätze besser nachgenutzt und die Repositorien besser vernetzt werden können, um semantische Recherchen zu ermöglichen.

Als Grundlage dienten nur die (linguistisch aufbereiteten) Dublin-Core-konformen OAI-Metadatenfelder *title*, *subject* und *description*,[65] keine Volltexte. Im Vorfeld wurden verschiedene Ansätze zur automatischen Klassifizierung getestet. Zwei dieser Ansätze stützten sich auf über Suchmaschinen bzw. die Wikipedia verfügbare Information, um semantische Zusammenhänge abzubilden, drei waren vektorraumbasiert, wovon einer SVMs und zwei die Methode der *Latent Semantic Analysis* (LSA) zur Anwendung brachten, einem patentgeschützten Verfahren, welches aus der Termhäufigkeitsverteilung die für ein Dokument repräsentativen Konzepte herauskondensiert.[66] In der Analyse erwies sich der SVM-Ansatz als der erfolgreichste und wurde daher weiterverfolgt.[67]

[62] Vgl. *Technische Informationsbibliothek analysiert Literatur mit Software von Averbis* (2011).
[63] Der volle Titel lautete: „Automatische Anreicherung von OAI-Metadaten mit Hilfe computerlinguistischer Verfahren und Entwicklung von Services für die inhaltsorientierte Vernetzung von Repositorien". Das Projekt war DFG-gefördert. Projektpartner waren die UB Bielefeld, das Text Technology Lab an der Universität Frankfurt und die Abteilung Automatische Sprachverarbeitung des Instituts für Informatik an der Universität Leipzig, siehe *Automatische Anreicherung von OAI-Metadaten* (o.D.).
[64] Siehe *Open Archives Initiative* (o.D.). „Die Open Archives Initiative (OAI) ist eine Initiative von Betreibern von Preprint- und anderen Dokumentenservern, um die auf diesen Servern abgelegten elektronischen Publikationen im Internet besser auffindbar und nutzbar zu machen. Dazu werden verschiedene einfache Techniken entwickelt und bereitgestellt, insbesondere das OAI Protocol for Metadata Harvesting (OAI-PMH) zum Einsammeln und Weiterverarbeiten von Metadaten. ... Das auf XML ... basierende OAI Protocol for Metadata Harvesting (OAI-PMH) dient ... zum Sammeln von Metadaten, die von so genannten Data Providern bereitgestellt werden. Die gesammelten Titeldatensätze werden dann von so genannten Service Providern aufbereitet und für Suchanfragen bereitgestellt. Aufgrund der Vielzahl von Metadatenformaten ist als kleinster gemeinsamer Nenner das Dublin-Core-Datenmodell vorgeschrieben", siehe *Wikipedia – Open Archives Initiative* (o.D.).
[65] Vgl. Waltinger et al. (2011, S. 33).
[66] Vgl. *Wikipedia – Latent semantic analysis* (o.D.).
[67] Vgl. Mehler & Waltinger (2009, S. 13–17).

Das im Projekt entwickelte Verfahren klassifiziert wissenschaftliche Dokumente automatisch bis in die dritte Ebene der DDC. Für das Training wurde ein Korpus von OAI-Metadaten aus der Datenbasis der *Bielefeld Academic Search Engine* (BASE)[68] für etwa 40 000 deutsche und 50 000 englische Texte aufgebaut.[69] Als Aufnahmebedingung musste für einen Datensatz mindestens eine DDC-Notation ermittelbar sein. Wenn ein Datensatz keine DDC-Notation enthielt, aber anderweitig klassifiziert worden war, so wurde sie anhand (manuell erstellter) Konkordanzen abgeleitet.[70] Die anschließend erstellten, nach der Vorkommenshäufigkeit eines Begriffs im Dokument und im Korpus gewichteten Merkmalsvektoren wurden benutzt, um für jede der ersten drei Ebenen der DDC und jede der Klassen auf diesen Ebenen eine separate SVM zu trainieren, die ihre Klasse von allen anderen auf der jeweiligen Ebene unterscheiden kann. Auf jeder Ebene konnten keine, eine oder mehrere Notationen vergeben werden.[71]

Die Qualität der Klassifizierung erreichte im Durchschnitt ein F-Maß von etwa 80%, auf der zweiten Ebene konnte ein F-Maß von 74% für deutsche und 63% für englische Texte erzielt werden, und auf der dritten ergab sich noch ein F-Maß von etwa 60% für beide Sprachen.[72] Problematisch war die recht ungleiche Verteilung der Trainingsdokumente auf die verschiedenen Klassen der DDC. Vor allem auf der dritten Ebene konnten einige Klassen nicht mit genügend Beispielen für hochqualitativ sacherschlossene Publikationen belegt werden und mussten ausgeschlossen werden, sodass nur 128 Klassen für das Deutsche und 88 für das Englische verblieben. Dies betraf vor allem Geisteswissenschaften, während etwa in der Physik aufgrund einer starken *Open-Access*-Tradition kein Mangel an Datensätzen zu elektronischen Publikationen herrscht und für diese sogar ein oberes Limit gesetzt werden musste, um den Trainingskorpus auszubalancieren.[73]

Durch das Projekt konnte die Anzahl der mit DDC-Notationen versehenen Dokumente im BASE-Index von ca. 400 000 auf über 1,7 Mio. gesteigert werden, für Nutzer ist nun auch ein Browsing entlang des DDC-Baumes möglich, und sowohl die angereicherten Metadaten als auch die entwickelten Klassifizierungsinstrumente können über entsprechende Schnittstellen nachgenutzt werden.[74]

[68] Siehe *Über BASE* (o.D.).
[69] Für Einzelheiten zur Erstellung des Korpus siehe Lösch et al. (2011).
[70] Welche Konkordanz jeweils zur Anwendung kommen musste, wurde in einem halbautomatischen Verfahren anhand der Struktur der Notationen ermittelt, siehe Lösch et al. (2011, S. 3).
[71] Vgl. Waltinger et al. (2011, S. 33–34).
[72] Die Genauigkeit, Vollständigkeit, und das F-Maß für einzelne Fächer schwankte jedoch stark von unter 10% bis über 90%, siehe Waltinger et al. (2011, S. 35–38).
[73] Vgl. Lösch et al. (2011, S. 5–6).
[74] Siehe *Automatische Anreicherung von OAI-Metadaten* (o.D., „Projektergebnisse“).

2.5 VZ Göttingen: Das Projekt Colibri (DDC)

Das auf ein Pica-Projekt zurückgehende VZG-Projekt „*Context generation and Linguistic tools for Bibliographic Retrieval Interfaces*“ (Colibri) befasst sich seit 2003 mit der Entwicklung eines Systems zur Bereitstellung von automatischen Verfahren zur DDC, insbesondere für die einheitliche und effiziente Inhaltserschließung und die Analyse und Synthese von DDC-Notationen zur Unterstützung von Anwendern der DDC.[75]

Die Klassifizierungskomponente *vc_dcl* kombiniert Verfahren aus dem *Information Retrieval* (Vektorprodukt) mit heuristischen Verfahren aus der künstlichen Intelligenz.[76] Als Testdaten dienten von der DNB bereitgestellte, DDC-haltige Titeldatensätze aus der Deutschen Nationalbibliographie. Die Klassifizierung stützt sich auf die aus diesen Datensätzen abgeleitete Wissensbasis *vc_DB*, erweitert um die (englischen) Fakten aus der Wissensbasis *vc_KB*. Eine Wissensbasis ist hier eine Menge von DDC-Klassen und eine DDC-Klasse (repräsentiert durch eine DDC-Notation) wird definiert durch eine Menge von Deskriptorwerten. Deskriptoren sind in diesem Fall Kategorien des Datenformats *Pica+*, deren Werte zur inhaltlichen Charakterisierung geeignet sind.[77]

Die Titeldatensätze werden als Tripel von DDC-Notationen,[78] Deskriptoren und Deskriptorwerten dargestellt. Zu jedem Deskriptorwert in einem zu klassifizierenden Datensatz werden alle DDC-Klassen herausgesucht, die diesen Deskriptorwert ebenfalls enthalten, und aus diesen über das Vektorprodukt als Ähnlichkeitsmaß die passendsten Notationskandidaten ermittelt. Hierbei gilt die Annahme, dass Deskriptorwerte, die in zu vielen DDC-Klassen auftreten, für die automatische Klassifizierung ungeeignet sind und daher mit Hilfe einer heuristischen Funktion ausgeschlossen werden müssen.[79]

Die Ergebnisse wurden sowohl intellektuell anhand einer sechsstufigen Skala als auch automatisch mittels eines stellenweisen Ziffernvergleichs von links nach rechts zwischen intellektuell und automatisch vergebener Notation bewertet.[80] Die automatische Bewertung ergab eine Übereinstimmung von ca. 65% in der Hauptklasse (also der ersten Stelle), ca. 50% in den ersten beiden und ca. 24% in

[75] Vgl. Reiner (2003, S. 3), Reiner (2009b, S. 3).
[76] Vgl. Reiner (2009a, Folie 15–17).
Mit der Entwicklung dieser Komponente wurde im Jahr 2006 begonnen, siehe Reiner (2010, S. 24).
[77] Vgl. Reiner (2009b, S. 4–5). Zu den verwendeten Pica+-Kategorien siehe Reiner (2009a, Folie 25).
[78] Hierbei wurden die Notationen in der Wissensbasis soweit gekürzt, dass sie in einer der DDC-Haupttafeln enthalten waren. Bei zu klassifizierenden Datensätzen wurde als Notation vor der Klassifizierung ein Dummy-Wert („XXX“) eingetragen, siehe Reiner (2009b, S. 5–6).
[79] Vgl. Reiner (2009b, S. 7–9, 12–13).
[80] Vgl. Reiner (2010, S. 26), Reiner (2009a, Folie 31).

den ersten drei Stellen.[81] Dabei wurden unter anderem signifikante Unterschiede zwischen den einzelnen DDC-Klassen und zwischen verschiedenen Publikationsformen (Reihen A, B und H der DNB) festgestellt.[82]

In das Verfahren wurden bis dahin keine Textteile, keine Lexika und insbesondere keine linguistischen Methoden miteingeschlossen.[83] Es gab Bestrebungen, die Klassifizierungsleistung mit Hilfe einer linguistischen Verarbeitung zu verbessern, jedoch bislang ohne nennenswerten Erfolg.[84] Eine weitere Bestrebung ist die Prüfung der Frage, ob und wie die ebenfalls im Projekt entwickelten Instrumente zur automatischen Synthese von DDC-Notationen auf der Basis ihrer Zerlegbarkeit in atomare Bestandteile zur weiteren Verbesserung der automatischen Klassifizierung eingesetzt werden können.[85]

2.6 ZB Zürich: Das Projekt ComSE (DDC)

Einem Grundsatzentscheid der Konferenz der Deutschschweizer Hochschulbibliotheken aus dem Jahr 2009 zufolge sollte der Personal- und Zeitaufwand für die intellektuelle Sacherschließung durch Fremddatenübernahme, Nutzung von Normdateien und insbesondere auch die Anwendung (halb-)automatisierter Erschließungsverfahren schrittweise reduziert werden. Als Reaktion darauf wurde an der Zentralbibliothek Zürich im Jahre 2011 ein Pilotprojekt gestartet, bei dem in Kooperation mit der IT-Dienstleistungsfirma Eurospider existierende Methoden des maschinellen Lernens und entsprechende Software auf ihre Anpassbarkeit an Sacherschließungsaufgaben geprüft werden sollten. Eigenentwicklungen waren zunächst nicht vorgesehen.[86]

Der im Rahmen des Projektes entwickelte *Digitale Assistent* ist ein System, welches auf ein Repositorium von Metadaten aus verschiedenen Quellen (Verbund, WorldCat, DNB) zurückgreift und computerunterstützte und intellektuelle Erschließungsprozesse miteinander verschränkt. Er bietet (neben statistischen Auswertungswerkzeugen) Unterstützung bei der Schlagwortvergabe und bei der Klassifizierung anhand der DDC. Datengrundlage für das Training der Klassifizierungssoftware waren von der DNB bereitgestellte Metadaten und Inhaltsverzeichnisse.

[81] Vgl. Reiner (2009b, S. 105).
[82] Vgl. Reiner (2009a, Folie 41–47).
[83] Vgl. Reiner (2009a, Folie 48).
[84] Getestet wurde die Software *Lingo*, vgl. *VZG-Verbundzentrale: Jahresbericht* (2012, S. 21).
[85] Vgl. Reiner (2008, Folie 20–23)
[86] Soweit nicht anders angegeben, beziehen sich die Ausführungen in diesem Abschnitt auf die Quellen Malits & Schäuble (2011), *Computerunterstützte Sacherschliessung: Der Digitale Assistent (Präsentation)* (2013) und auf das interne Dokument Malits & Schäuble (2013), welches der Autorin von der Projektkoordinatorin Andrea Malits per Email zugesandt wurde.

Seit Mitte Oktober 2013 wird von den im Projekt vertretenen Fachreferenten mit dem Digitalen Assistenten produktiv gearbeitet.[87] Im Anschluss an die maschinelle Klassifizierung wird jedoch weiterhin eine Überprüfung und eine feinere Einteilung in die Fachgebiete nach hauseigenem System durch die Fachreferenten vorgenommen. Auch hier zeichnete sich ab, dass sich gewisse Fachgebiete weniger gut für eine automatische Erschließung eignen als andere. Weitere Ziele sind die Berücksichtigung französischer und englischer Dokumente und die Einbeziehung von Volltexten, dafür wird mit Verlagen über die Bereitstellung der benötigten elektronischen Daten verhandelt.

2.7 Weitere Projekte

Es gibt zahlreiche weitere Projekte zur Automatisierung von Klassifizierungsverfahren. In diesem Abschnitt streifen wir noch einige Beispiele aus zwei unterschiedlichen Kategorien von weiteren Projekten, zum einen solche, die sich auf ein einzelnes Fachgebiet beschränken und zum anderen solche, die im Rahmen von theoretischen Forschungsstudien oder von Abschlussarbeiten entstanden sind.

An der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW) wurden ab 2009 verschiedene Softwarelösungen für die maschinelle Indexierung evaluiert mit dem Ziel, geeignete automatische Komponenten in die Sacherschließung hausinterner und externer digitaler Dokumente einzupassen. Die Entscheidung fiel auf das Produkt *Decisiv Categorization* der Firma Recommind, ein rein auf Wortmusterhäufigkeit basierendes, statistisches Verfahren.[88] Durch eine Einordnung in den Standard-Thesaurus Wirtschaft wurde der Ansatz zu einer begriffsorientierten semantischen Methode weiterentwickelt („*Additionsverfahren*"). Output und Lernfähigkeit des Systems werden zusätzlich optimiert durch die Einbettung in einen semi-automatisierten Ablauf und damit durch eine ständige intellektuelle Verbesserung der Trainingsbasis.[89]

Auswertungen zeigten, dass die Konsistenz zwischen intellektueller und automatischer Indexierung stark schwankte und insgesamt nur 36% erreichte. Außerdem verwendeten menschliche Sacherschließer 71% der zur Verfügung stehenden Begriffe, die Automatik jedoch nur 29%, was auf eine mangelnde Abdeckung durch die recht kleine Trainingsbasis zurückgeführt werden kann, aber eventuell

[87] Persönliche Mitteilung per Email von der Projektkoordinatorin Andrea Malits.
[88] „Die ... zur automatischen Indexierung eingesetzte Software „Decisiv Categorization" basiert auf der vom Hersteller patentierten CORE-Technologie und bedient sich der Probabilistic Latent Semantic Analysis (PLSA)", siehe Groß & Faden (2010, S. 1126). Siehe auch *Core Platform* (o.D.).
[89] Vgl. Groß & Faden (2010, S. 1127–1128).

auch auf die mangelnde Trennschärfe zwischen den Begriffen im Thesaurus. Weitere Schritte sind die Erweiterung dieser Basis um von Verlagen zur Verfügung gestellte Daten und das Training des Systems für eine Erschließung mit der hauseigenen Standardklassifikation.[90]

Ein weiteres fachspezifisches Projekt befasste sich mit der Einsetzbarkeit der Indexierungssoftware AUTINDEX für die PSYNDEX-Datenbank am Zentrum für Psychologische Information und Dokumentation. Die Software kombiniert linguistische und statistische Ansätze und nutzt ebenfalls einen Thesaurus. Eine Evaluation aus dem Jahre 2011 lieferte ähnliche Ergebnisse wie die beim vorigen Projekt genannten.[91]

Eine klassische Referenz für eine theoretische Untersuchung und entsprechende Experimente zur automatischen DDC-Klassifizierung mit Methoden des maschinellen Lernens ist Wang (2009). Als Lösung für die ungleiche Verteilung von Trainingsdokumenten über die verschiedenen Klassen und Tiefen der DDC schlägt Wang eine Verflachung und Kondensierung des DDC-Hierarchiebaumes vor. Der von Wang präsentierte interaktive Algorithmus für SVM-basiertes Klassifizieren auf der Grundlage eines derart umstrukturierten Baumes erreicht für beliebig spezifische DDC-Klassen eine Präzision von über 90%, benötigt für dieses Ergebnis jedoch bis zu drei Mal eine intellektuelle Auswahlentscheidung (aus bis zu fünf Kandidaten) durch einen menschlichen Experten.[92]

Eine weitere theoretische Untersuchung von Joorabchi & Mahdi (2011) nutzt die in Zitationsdatenbanken wie *CiteSeer* verzeichnete Verweisungsstruktur zwischen Dokumenten aus und ermittelt auf Grund der in den Metadaten hinterlegten Klassifizierungen für die Referenzen sowohl im als auch auf das zu klassifizierende Dokument automatisch (mit Hilfe eines Gewichtungsmechanismus') die wahrscheinlichste DDC-Klasse für dieses Dokument. Hier findet also kein Training mit intellektuell erschlossenen Daten statt, deswegen bezeichnet man einen solchen Ansatz auch als *unüberwacht*.

[90] Vgl. Groß & Faden (2010, S. 1131, 1133–1135).
[91] Vgl. Gerards et al. (2006), Gerards (2011).
[92] Vgl. Wang (2009, S. 2280).

Auch der in der BA-Arbeit von Sommer (2012) beschriebene Ansatz zur Klassifizierung von Hochschulschriften kommt ohne Training aus, sondern setzt (wie das oben genannte Projekt aus dem Bereich Wirtschaft) auf ein begriffsorientiertes Verfahren. Zur Abbildung von semantischen Begriffsrelationen wird die im DFG-Projekt CrissCross[93] von der DNB erstellte Semantic-Web-Ontologie eingebunden, welche die Einträge der Schlagwortnormdatei (SWD) mit DDC-Notationen verknüpft.[94] Der Klassifizierer basiert auf der Open-Source-Software *GATE*, in die verschiedene linguistische Werkzeuge miteingebunden und über eine weitere Schnittstelle auf die angereicherte SWD-Ontologie zugegriffen werden kann. In Anlehnung an das Projekt der UB Bielefeld (Abschnitt 2.4) werden als Datengrundlage OAI-Metadaten herangezogen, jedoch nur die Felder *title*, *subject* und *description*. Im Test befand sich in 80% der Fälle die korrekte Lösung unter den ersten drei vorgeschlagenen Notationen aus der DDC.[95]

Ebenfalls auf die SWD bezieht sich die BA-Arbeit von Probstmeyer (2009). In dieser Arbeit werden automatisch generierte Korrelationen zwischen der SWD und der RVK ausgewertet und unter anderem auf ihre Einsatzmöglichkeiten in der computergestützten Sacherschließung hin untersucht.

In ihrer Masterarbeit entwickelt Helmbrecht-Schaar (2007) einen Prototypen für die halbautomatisierte Klassifizierung von Online-Texten aus der Informatik anhand der fachspezifischen CR-Klassifikation. Als Basis dient das Open-Source-Produkt *MyCoRe*, welches diverse Funktionalitäten von der Speicherung und Suche von Metadaten und Volltexten über Editoren bis hin zu einem hierarchischen Klassifizierungssystem bereitstellt. Die gewählte Klassifikation wird dahingehend erweitert, dass ihre Klassen in einem Lernverfahren um Synonyme und weitere Deskriptoren angereichert werden können. Die relevanten Terme werden mit Hilfe eines statistisch-linguistischen Werkzeugs aus dem zu klassifizierenden Dokument extrahiert, für eine Suche in den Deskriptoren der CR-Klassen zur Ermittlung der geeigneten Klasse verwendet und nach Abschluss des Vorgangs ebenfalls (unter intellektueller Kontrolle, über eine Nutzeroberfläche) zur Beschreibung der resultierenden Klasse in das System eingespeist.[96]

Abschließend sei die Diplomarbeit von Wille (2006) genannt, in deren Rahmen der Autor mehrere Klassifizierungsalgorithmen aus dem Bereich des maschinellen Lernens testet und vergleicht. Zu diesem Zweck wurde aufbauend auf ei-

[93] Siehe *CrissCross* (o.D.).
[94] Vgl. Sommer (2012, S. 36).
[95] Vgl. Sommer (2012, S. 53).
[96] Vgl. Helmbrecht-Schaar (2007, S. 57–58).

nem existierenden Perl-Modul eine Programmumgebung geschaffen, welche flexible Schnittstellen zum Datenhaltungssystem und zur Einbindung einer Indexierungssoftware (wie z.B. Lingo) bietet und eine Reihe verschiedener Klassifizierungsalgorithmen unterstützt. Getestet wurde unter anderem die Methode der SVMs, *k nearest neighbour* und *Decision Tree*. Auch hier zählten SVMs (zusammen mit dem *Decision-Tree*-Algorithmus) nach einer Trainingsphase mit intellektuellen Daten zu den erfolgreichsten Klassifikatoren.[97]

3. Analyse und Fazit

Es gibt viele Kriterien, nach denen Projekte zur automatischen Klassifizierung angeordnet werden können, und wir erheben für den folgenden Versuch keinen Anspruch auf Vollständigkeit. Einige Einteilungen sind eventuell eher kontinuierliche Skalen:

- *Chronologische Anordnung.* Eine solche Darstellung könnte sinnvoll sein vor dem Hintergrund der jeweils aktuellen technologischen Entwicklung.

- *Halb- vs. vollautomatische Verfahren.* Bei den betrachteten Projekten wurde in der Mehrzahl der Fälle ein vollautomatisches Verfahren entwickelt, dessen Ergebnis dann jedoch einer intellektuellen Kontrolle unterzogen wurde. Einzig der Ansatz von Wang (2009) benötigt intellektuellen Input durch einen Experten während des Klassifizierungsprozesses selbst.

- Nach *Art der Datengrundlage*, z.B. dem *verwendeten Textanteil*:
  - einzelne Phrasen, Stich- und Schlagwörter aus den Metadaten (wie z.B. im Projekt der UB Mannheim oder dem der VZ Göttingen; siehe Abschnitte 2.1, 2.5)
  - Abstracts/Inhaltsverzeichnisse (UB Bielefeld, ZB Zürich; Abschnitte 2.4, 2.6)
  - Volltexte (wie z.B. im Projekt PETRUS der DNB; Abschnitt 2.2)

  wobei Projekte, die größere Textanteile nutzen, in der Regel auch die kürzeren Informationen auswerten. Ein weiteres, technisches Kriterium ist das *Datenformat der verwendeten Daten*, z.B. XML, MAB2 (UB Mannheim) oder Pica+ (VZG).

[97] Vgl. Wille (2006, S. 10–11, 22–23, 26–27, 29–30).

- Nach *Publikationsform*: Originär elektronische Publikationen vs. Digitalisierungen von Druckausgaben (über den Sinn dieser Gegenüberstellung lässt sich jedoch streiten, da mit der Durchsuchbarkeit von digitalisierten Texten mittels OCR die Voraussetzung für eine automatisierte Erschließung in beiden Fällen gegeben ist), Monographien, Hochschulschriften, Texte aus einem eingeschränkten Fachgebiet (siehe hierzu vor allem die Projekte aus Abschnitt 2.7).

- Nach der *gewählten Klassifizierungsmethode*. Hier schlagen wir zunächst eine Einteilung der Methoden nach einer Art Dreistufenmodell vor:
  - Direktübernahme einer Notation der Zielklassifikation aus einer fremden Quelle (also quasi eine Nullleistung, was die Klassifizierung anbetrifft)
  - Mittelbare Übernahme durch eine Abbildung irgendeiner Art. Hierzu kann eventuell die Arbeit von Joorabchi & Mahdi (2011) gezählt werden, in der die Einbindung eines Dokuments in ein Netz von Zitationen ausgenutzt wird, und auch die beiden BA-Arbeiten aus Abschnitt 2.7, bei denen mit Hilfe von Deskriptoren eine Abbildung aus einem zugrundeliegenden semantischen Netz (einer Normdatei) in die Zielklassifikation geschaffen wird.
  - („Echte“) automatisierte Klassifizierung mit komplexeren Methoden, und hier wiederum weitere Einteilungen nach diversen Kriterien, beispielsweise:
    * Klassifizierungsverfahren mit (z.B. das Projekt der UB Mannheim) und ohne (VZG) linguistische Verarbeitungsschritte
    * Klassifizierungsverfahren mit (z.B. das Projekt der ZBW; Abschnitt 2.7) und ohne (UB Mannheim) statistische Komponenten
    * Verfahren aus der Künstlichen Intelligenz, insbesondere aus dem Bereich des maschinellen Lernens wie das fallbasierte Schließen (UB Mannheim) oder SVMs (DNB, TIB, UB Bielefeld), und Verfahren aus dem Information Retrieval wie z.B. die Anwendung des Vektorprodukts (VZG).

Der in Abschnitt 2. unternommene Querschnitt durch die aktuelle Projektlandschaft im deutschsprachigen Raum erlaubt folgende Feststellungen: Eine Reihe der bedeutenden, fachübergreifenden Projekte zur automatischen Klassifizierung

stützt sich auf maschinelle Lernverfahren, insbesondere auf die Methode der SVMs. Viele Projekte binden auch eine linguistische Vorverarbeitung mit ein. Es werden vermehrt kommerzielle Produkte übernommen und angepasst, etwa die Averbis Extraction Platform oder Software der Firma Recommind, oder (für die Vorverarbeitung) die Indexierungssoftware Lingo.

Das Training maschineller Klassifizierungsinstrumente ist allerdings recht zeit- und datenintensiv und beinhaltet entsprechende Herausforderungen:

- Die Trainingsdaten müssen beschafft werden. Viele Projekte führen unzufriedenstellende Ergebnisse auf zu kleine Trainingsbasen zurück. Verhandlungen mit Verlagen mit dem Ziel, Metadatensätze zu bereits lizensierten Titeln in elektronischer Form bereitgestellt zu bekommen, scheinen sich häufig schwierig zu gestalten.

- Die Trainingsdaten müssen eine entsprechende Qualität aufweisen. Die Erschließung muss korrekt, vollständig, möglichst tief und möglichst konsistent sein, rein formal definierte Notationen („Festschrift“ o.Ä.) müssen ausgeschlossen werden.

- Die Trainingsdaten sollten möglichst gleichmäßig über die zu vergebenden Klassen verteilt sein, damit für jede Klasse genügend repräsentative Beispiele vorhanden sind. Dies ist für fachübergreifende Projekte meist besonders problematisch: In manchen Fächern (v.a. Naturwissenschaften) herrscht eine Fachkultur, die eine Fülle frei zugänglicher elektronischer Publikationen samt zugehöriger Metadaten begünstigt, in den Geisteswissenschaften ist das jedoch noch nicht der Fall.

Die Aufgabe der Klassifizierung wird erleichtert durch eine sprachliche Homogenität der Trainingsdokumente – sowohl, was die natürliche (Deutsch, Englisch), als auch, was die Fachsprache betrifft. Verschiedene Fächer weisen unterschiedliche Textmuster und unterschiedliche Terminologien auf. Manche Fachterminologien sind eventuell intrinsisch unscharf, sodass die einzelnen Begriffe nur schwer voneinander abzugrenzen sind. Solche Fächer (auch hier tendenziell die Geisteswissenschaften) stellen für die Klassifizierung naturgemäß eine größere Herausforderung dar. Die Konsistenz der Klassifizierungsergebnisse kann möglicherweise erhöht werden durch das Einbeziehen von Normdateien oder anderen semantischen Netzen zu den vorkommenden Themenkomplexen. Ein weiterer begünstigender Faktor neben scharf trennbaren Klassen (horizontale Unterscheidbarkeit) in einer Klassifikation ist eine saubere hierarchische Struktur zur Abbildung der

Ober-/Unterklassenbeziehung.[98] Eine Verminderung der Anzahl der möglichen Klassen (etwa durch eine Beschränkung auf die oberste Ebene der DDC) verringert die Komplexität der Aufgabe auf triviale Weise ebenfalls, macht aber auch das Ergebnis weniger wertvoll, da weniger spezifisch.

Die Erfolgsquote aktueller automatischer Klassifizierungsvorhaben bewegt sich (gemessen mit diversen Bewertungsmaßen, hauptsächlich dem F-Maß) ganz grob im Raum von 60% bis 85%. Die Schlussfolgerungen in den Projekten reichen von „Automatische Klassifizierungen können die intellektuelle Sacherschließung nicht ersetzen“ [99] bis hin zu der vergleichsweise optimistischen Aussage, dass ein Einsatz im laufenden Betrieb nur begleitet von einem kontinuierlichen Qualitätsmanagement erfolgen kann.[100]

Es gibt weitere Ansätze, den Aufwand für inhaltliche Sacherschließung zu reduzieren und dennoch die Flexibilität nachzubilden, mit der ein menschlicher Klassifizierer anhand seines Weltwissens entscheidet, was inhaltlich relevante Information ist und was nicht – insbesondere dann, wenn diese Information an unerwarteten Stellen zu finden ist, sodass eine Abbildung in ein felderbasiertes Datenmodell ein anspruchsvolle Aufgabe darstellt. Denkbar sind etwa Kombinationen mit heuristischen Methoden[101] oder auch mit dem sogenannten *social tagging*, bei dem die intellektuelle Erschließungsleistung von einer großen Zahl von Nutzern eingebracht wird.[102]

Wenn ein automatisiertes Verfahren sowohl im Vorlauf bei der Erstellung von Trainingsdaten als auch im Nachgang zur Kontrolle intellektuellen Inputs bedarf, so muss das Verhältnis zwischen menschlichem Aufwand und der Menge schlussendlich erfolgreich klassifizierter Dokumente hinreichend klein sein, um es noch attraktiv zu machen. Für Bestände, die aufgrund ihrer Fülle anson-

[98] Aus diesem Grunde finden beispielsweise an der Bibliothek der Universität Konstanz zur Zeit umfangreiche Projektarbeiten zur Bereinigung der hauseigenen Systematik statt, mit dem Fernziel eines Klassifizierungsvorschlagstools für den gehobenen Dienst, einer verbesserten Austauschbarkeit von Sacherschließungsleistungen mit anderen Institutionen und einer Anbindung an die GND mit Hilfe von Semantic-Web-Technologien, siehe Kasprzik (2013, S. 4–5).

[99] Vgl. Mensing (2011, Folie 15). Siehe auch Pfeffer (2007a, S. 30): „Für eine vollautomatische systematische Erschließung ohne Kontrolle durch einen Experten ... ist der Anteil der schlecht oder falsch klassifizierten Titel noch zu hoch“ oder Reiner (2010, S. 28): „Die inspizierten automatischen (DDC-)Klassifizierer arbeiten besser als der Zufall, aber für einen professionellen umfangreichen Einsatz für Mio. von zu klassifizierenden ... Titeldatensätzen sind sie noch nicht geeignet“.

[100] Ein solches Qualitätsmanagement kann immerhin ebenfalls mit automatischen Methoden geschehen, siehe Mödden & Tomanek (2012, S. 23–24): „Durch einen kontinuierlichen maschinellen Vergleich der Sachgruppen aus maschineller und intellektueller Erschließung soll die Qualität der maschinellen Prozesse fortlaufend beobachtet werden, um bei Bedarf nachzusteuern“.

[101] Vgl. z.B. auch Pfeffer (2007a, S. 26–27, 30).

[102] Siehe dazu z.B. Zubiaga et al. (2011). Die Autoren kombinieren *social tagging* mit SVMs und wählen als Zielklassifikationen die DDC und die *Library of Congress Classification* (LCC).

sten gar nicht erschlossen würden, ist eine automatische Klassifizierung in jedem Fall eine gute Option. Dies betrifft vor allem die ständig anwachsende Masse der reinen Online-Publikationen, die sich andererseits mit ihrer elektronischen Verfügbarkeit für computerbasierte Methoden ja auch anbieten.[103] Überdies könnte man sich angesichts der genannten Erfolgswerte fragen, wie hoch denn die Übereinstimmung zwischen den durch verschiedene menschliche Experten oder auch den vom selben Experten an verschiedenen Tagen vergebenen Notationen ist und den Pessimismus weiter relativieren, da automatisch erzeugte Klassifizierungen den Vorteil haben, aufgrund ihrer mechanischen Entstehungsweise in sich konsistent(er) zu sein.[104]

[103] Ob sich die für die maschinelle Erschließung problematische Knappheit digital verfügbarer Dokumente und Metadaten aus dem Bereich der Geisteswissenschaften im Zuge der (Retro-) Digitalisierung oder auch aufgrund sich wandelnder Fachgepflogenheiten langfristig entschärft, bleibt abzusehen.

[104] Da die meisten hier untersuchten Verfahren sich auf intellektuell erstellte Daten stützen, könnte man weiter fragen, inwieweit solche Inkonsistenzen schon in die Trainingsdaten miteingeflossen sind und damit den Lernprozess beeinflusst haben.

Literatur

Alle Internetquellen wurden zuletzt abgerufen am 22.11.2013.

Automatische Anreicherung von OAI-Metadaten
http://www.ub.uni-bielefeld.de/wiki/AutoOAI

Averbis Textanalyse http://www.averbis.de/de/technologies/text_analytics

Bähr, T. (2010). LINSearch: Chancen und Risiken im Grenzbereich zwischen intellektueller Erschließung und automatischer Klassifizierung (Präsentation).
http://verbundkonferenz.gbv.de/wp-content/uploads/2010/09/TIB.pdf

Computerunterstützte Sacherschliessung: Der Digitale Assistent (Präsentation). (2013).
http://www.eurospider.ch/fileadmin/pdf/2013_05_29_ComSE_Bern_v3sel.pdf

Core Platform http://www.recommind.de/core-platform

CrissCross http://linux2.fbi.fh-koeln.de/crisscross/index.html

Dewey-Dezimalklassifikation
http://www.ddc-deutsch.de/Subsites/ddcdeutsch/DE/Home/home_node.html

Gerards, M. (2011). Semiautomatische Erschließung von Psychologie-Information (Präsentation).
http://www.dnb.de/SharedDocs/Downloads/DE/DNB/wir/petrus/gerardsSemiautomatischeErschliessungZpid.pdf

Gerards, M., Gerards, A. & Weiland, P. (2006). Der Einsatz der automatischen Indexierungssoftware AUTINDEX im Zentrum für Psychologische Information und Dokumentation (ZPID).
http://www.zpid.de/download/PSYNDEXmaterial/autindex.pdf

Gödert, W., Lepsky, K. & Nagelschmidt, M. (2012). *Informationserschließung und automatisches Indexieren: Ein Lehr- und Arbeitsbuch*. X.media.press. Berlin u.a.: Springer.

Groß, T. & Faden, M. (2010). Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. *Bibliotheksdienst*, *44*(12), 1120–1135.

Helmbrecht-Schaar, A. (2007). Entwicklung eines Verfahrens der automatischen Klassifizierung für Textdokumente aus dem Fachbereich Informatik mithilfe eines fachspezifischen Klassifikationssystems.
http://www.ib.hu-berlin.de/ ~ kumlau/handreichungen/h200/h200.pdf

Joorabchi, A. & Mahdi, A. E. (2011). An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. *Journal of Information Science*, *37*(5), 499–514.

Kasprzik, A. (2013). Projektbericht: Implementierung eines Hierarchisierungsalgorithmus' für die Konstanzer Systematik. http://nbn-resolving.de/urn:nbn:de:bsz:352-241667

LINSearch: Linguistisches Indexieren und Suchen
http://www.tib-hannover.de/en/research-and-development/finished-projects/linsearch/

Lösch, M., Waltinger, U., Horstmann, W. & Mehler, A. (2011). Building a DDC-annotated corpus from OAI metadata. *Journal of Digital Information*, *12*(2).

Malits, A. & Schäuble, P. (2011). Computerunterstützte Sacherschliessung: Pilotprojekt. http://www.eurospider.com/fileadmin/pdf/Projektinfo.pdf

Malits, A. & Schäuble, P. (2013). Ein Assistent für die Beschlagwortung: Das Pilotsystem für die computerunterstützte Sacherschliessung der Zentralbibliothek Zürich (internes Dokument).

Mehler, A. & Waltinger, U. (2009). 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*, *27*(4), 520–539.

Mensing, P. (2011). Automatische Klassifizierung in der TIB (Präsentation). https://www.gbv.de/cls-download/fag-erschliessung-und-informationsvermittlung/arbeitsdokumente-fag-ei/vorstellung-der-automatischen-klassifizierung-an-der-tib

Mödden, E. & Tomanek, K. (2012). Maschinelle Sachgruppenvergabe für Netzpublikationen. *Dialog mit Bibliotheken*, *25*(1), 17–24.

Oberhauser, O. (2005). *Automatisches Klassifizieren: Entwicklungsstand – Methodik – Anwendungsbereiche*. Europäische Hochschulschriften. Peter Lang.

Open Archives Initiative http://www.openarchives.org/

PETRUS - Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek http://www.dnb.de/DE/Wir/Projekte/Abgeschlossen/petrus.html

Pfeffer, M. Self-Classification. http://blog.bib.uni-mannheim.de/Classification/?page_id=2

Pfeffer, M. (2007a). *Automatische Vergabe von RVK-Notationen anhand von bibliografischen Daten mittels fallbasiertem Schließen* (Masterarbeit, im Rahmen des postgradualen Fernstudiums Master of Arts (Library & Information Science)). http://blog.bib.uni-mannheim.de/Classification/wp-content/uploads/2007/10/main.pdf

Pfeffer, M. (2007b). Automatische Vergabe von RVK-Notationen (Präsentation). http://blog.bib.uni-mannheim.de/Classification/wp-content/uploads/2007/10/hu-berlin-2007-2.pdf

Pfeffer, M. (2008). Automatische Vergabe von RVK-Notationen mittels fallbasiertem Schließen (Präsentation). http://blog.bib.uni-mannheim.de/Classification/wp-content/uploads/2008/6/bibtag2008_rvk.pdf

Probstmeyer, J. (2009). *Analyse von maschinell generierten Korrelationen zwischen der Regensburger Verbundklassifikation (RVK) und der Schlagwortnormdatei (SWD)* (Bachelorarbeit, Hochschule der Medien Stuttgart, Fakultät Information und Kommunikation). http://nbn-resolving.de/urn:nbn:de:bsz:900-opus-6670

Recommind CORE Platform http://www.recommind.de/core-platform

Regensburger Verbundklassifikation http://rvk.uni-regensburg.de/

Reiner, U. (2003). VZG-Projekt Colibri: Überblick, Stand, Ergebnisse. Juli–Dezember 2003. https://www.gbv.de/cls-download/fag-erschliessung-und-informationsvermittlung/vzg-dokumente/vzg-projekt-colibri-stand-12-2003/colibri01-04-03-11-without-appendix.pdf/view

Reiner, U. (2008). Computer-aided assignment of DDC numbers (Präsentation). http://www.qucosa.de/fileadmin/data/qucosa/documents/5668/data/reiner.pdf

Reiner, U. (2009a). Automatische DDC-Klassifizierung von bibliografischen Titeldatensätzen (Präsentation). 98. Deutscher Bibliothekartag: Ein neuer Blick auf Bibliotheken. http://www.opus-bayern.de/bib-info/volltexte/2009/736/

Reiner, U. (2009b). VZG-Projekt Colibri: Bewertung von automatisch DDC-klassifizierten Titeldatensätzen der Deutschen Nationalbibliothek (DNB). August 2008 – Februar 2009. http://taipan.dyndns.org/ ~ ul/colibri05.pdf

Reiner, U. (2010). Automatische DDC-Klassifizierung. *Dialog mit Bibliotheken*, *21*(1), 23–29.

Schöning-Walter, C. (2011). Automatische Erschließungsverfahren für Netzpublikationen. *Dialog mit Bibliotheken*, *23*(1), 31–36.

Siegmüller, R. (2007). Verfahren der automatischen Indexierung in bibliotheksbezogenen Anwendungen. http://www.ib.hu-berlin.de/ ~ kumlau/handreichungen/h214/h214.pdf

Sommer, M. (2012). *Automatische Generierung von DDC-Notationen für Hochschulveröffentlichungen* (Bachelorarbeit, Hochschule Hannover, Fakultät III – Medien, Information und Design). http://opus.bsz-bw.de/fhhv/volltexte/2012/397/

Technische Informationsbibliothek analysiert Literatur mit Software von Averbis. (2011). http://www.pressebox.de/inaktiv/averbis-gmbh/Technische-Informationsbibliothek-analysiert-Literatur-mit-Software-von-Averbis/boxid/468358

Über BASE http://www.base-search.net/about/de/index.php

VZG-Verbundzentrale: Jahresbericht. (2012). http://www.gbv.de/Verbundzentrale/Publikationen/PDF/JB2012_mittel.pdf

Waltinger, U., Mehler, A., Lösch, M. & Horstmann, W. (2011). Hierarchical classification of OAI metadata using the DDC taxonomy. (Bd. 6699, S. 29–40). Advanced Language Technologies for Digital Libraries. Springer.

Wang, J. (2009). An extensive study on automated Dewey decimal classification. *Journal of the American Society for Information Science and Technology*, *60*(11), 2269–2286.

Was lange währt… : Automatische Fächerklassifizierung in GetInfo über die Facette “Fach”. (2012). http://blogs.tib-hannover.de/tib/2012/12/18/was-lange-waehrt-automatische-faecherklassifizierung-in-getinfo-ueber-die-facette-fach/

Wikipedia – Beurteilung eines Klassifikators http://de.wikipedia.org/wiki/Beurteilung_eines_Klassifikators

Wikipedia – Indexierung http://de.wikipedia.org/wiki/Indexierung

Wikipedia – Latent semantic analysis http://de.wikipedia.org/wiki/Latent_Semantic_Analysis

Wikipedia – Open Archives Initiative http://de.wikipedia.org/wiki/Open_Archives_Initiative

Wille, J. (2006). *Automatisches Klassifizieren bibliographischer Beschreibungsdaten: Vorgehensweise und Ergebnisse* (Diplomarbeit, Fachhochschule Köln, Fakultät für Informations- und Kommunikationswissenschaften). http://hdl.handle.net/10760/7790

Zubiaga, A., Körner, C. & Strohmaier, M. (2011). Tags vs shelves: From social tagging to social classification. In *Proceedings of the 22nd ACM conference on hypertext and hypermedia* (S. 93–102). ACM. http://doi.acm.org/10.1145/1995966.1995981