Christa Schöning-Walter

Automatische Erschließungsverfahren für Netzpublikationen

Zum Stand der Arbeiten im Projekt PETRUS

Die Deutsche Nationalbibliothek (DNB) hat damit begonnen, ihre Erschließungsprozesse zu automatisieren, um die Publikationen in ihrem Bestand und die bibliografischen Metadaten trotz der enorm anwachsenden Sammlung gedruckter und digitaler Medien so schnell wie möglich für die Nutzung zur Verfügung zu stellen.

Den organisatorischen Rahmen bildet das Projekt PETRUS.[1)] Dieses Akronym steht für »Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek«. Existierende Softwarelösungen aus den Bereichen Datenanalyse, Text-Mining und Information Retrieval sollen genutzt werden, um die Metadaten für die Nationalbibliografie direkt aus schon vorhandenen oder mitgelieferten Titeldaten sowie aus maschinenlesbaren Volltexten, Inhaltsverzeichnissen, Abstracts, Klappentexten o. ä. zu generieren. Als Anwendungsfall stehen während des Projektes primär die monografischen Netzpublikationen im Blickpunkt der automatischen Erschließung. Über das Projekt hinaus soll aufgezeigt werden, wie auch weitere Medien – beispielsweise bestimmte gedruckte Publikationen oder die Artikel wissenschaftlicher Zeitschriften – mit maschineller Unterstützung formal und inhaltlich erschlossen werden können. Ziel ist es, die intellektuelle Erschließung zu entlasten und die sich bietenden technischen Möglichkeiten auch für den Nachweis von Einzelbeiträgen in der Nationalbibliografie zu nutzen.

Monografische Netzpublikationen stehen im Fokus der automatischen Erschließung

Anfang 2010 hat sich die DNB entschieden, die steigende Zahl digitaler Publikationen grundsätzlich nicht mehr intellektuell zu bearbeiten. Vielmehr soll das Projekt PETRUS dazu führen, dass nicht nur Import, Archivierung und Bereitstellung, sondern auch formale und inhaltliche Erschließung von Netzpublikationen als weitgehend automatische Geschäftsprozesse ablaufen. Notwendigerweise müssen damit auch neue, geeignete Qualitätskriterien definiert werden. Die Prozessabläufe und Anforderungen an die Gütemerkmale der Metadaten werden voraussichtlich in Abhängigkeit vom Publikationstyp variieren, je nachdem, ob beispielsweise Monografien eines wissenschaftlichen Verlages, die Schriftensammlung von einem Hochschulserver, Book-on-Demand-Veröffentlichungen oder Ausschnitte aus dem Internet bearbeitet werden.

Definition neuer Qualitätskriterien

Die geplanten automatischen Erschließungsprozesse sollen in aufeinander aufbauenden Bearbeitungsstufen zu einer schrittweisen Anreicherung der Titeldaten führen. Die Herkunft der Metadaten und bestimmte Qualitätsmerkmale werden im Datensatz vermerkt, sodass auch eine eventuell eingeschränkte Vertrauenswürdigkeit erkennbar ist. Damit bleibt nachvollziehbar, ob die formalen und inhaltlichen Beschreibungen in der Nationalbibliografie nach bibliothekarischen Standards intellektuell erstellt, mit automatischen Verfahren erzeugt oder aus anderen Quellen übernommen wurden. Die Leistungsfähigkeit automatischer Erschließungssysteme wird voraussichtlich mit der technologischen Entwicklung steigen, sodass zu einem späteren Zeitpunkt bei Bedarf eine erneute maschinelle Bearbeitung der Metadaten durchgeführt werden kann. Nur in Ausnahmefällen soll eine intellektuelle Nachbesserung erfolgen. Aufwände zur Verbesserung der Erschließungsergebnisse sollen vorrangig in die Steuerung und Kontrolle der automatischen Verfahren sowie in die Pflege und Weiterentwicklung der Normdateien investiert werden.

Modulare Bearbeitungsstufen zur Anreicherung von Titeldaten

Die Normdateien behalten auch im Umfeld der automatischen Erschließungsprozesse ihre herausragende Bedeutung und werden als Erschließungsstandards in die maschinellen Verfahren mit eingebunden. Die DNB setzt derzeit die Schlagwortnormdatei (SWD), die Personennamendatei (PND) und die Gemeinsame Körperschaftsdatei (GKD) als normiertes Vokabular für ihre Erschließung ein und ist in Kooperation mit Bibliotheksverbünden

Herausragende Bedeutung von Normdateien

und anderen Institutionen aus Deutschland, Österreich und der Schweiz auch an deren fortlaufender Aktualisierung und Erweiterung maßgeblich beteiligt. Derzeit werden PND, SWD und GKD sowie die Einheitssachtiteldatei (EST) des Deutschen Musikarchivs zur Gemeinsamen Normdatei (GND) zusammengeführt. Bestehende Formatunterschiede, die parallele Haltung von Datensätzen in mehreren Normdateien, beispielsweise für Körperschaften und Geografika, sowie unterschiedliche Ansetzungsregeln für die Formal- und Inhaltserschließung sollen damit überwunden werden. Die GND wird voraussichtlich Ende 2011 in den Produktivbetrieb übernommen werden.

Im Jahr 2010 war die Erprobung ausgewählter automatischer Erschließungsverfahren und -methoden der wichtigste Arbeitsschwerpunkt im Projekt PETRUS. Die Untersuchungen konzentrierten sich dabei auf die vier zu Projektbeginn definierten Anwendungsszenarien:

- automatische Erkennung paralleler oder ähnlicher Ausgaben von Publikationen und Austausch von Metadaten zwischen den Titeldaten,
- automatische Generierung von Datensätzen in der PND beim Import neuer Titel und Erstellung von Relationen zwischen Personennamen und Titeldaten,
- automatische Einordnung der Netzpublikationen in die Systematik der DDC-Sachgruppen sowie
- automatische Vergabe von Schlagwörtern auf Grundlage des kontrollierten Vokabulars der SWD.

Testlizenzen für vier Softwareprodukte in der Erprobung

Für Experimente zur Untersuchung der Machbarkeit wurden Testlizenzen für vier verschiedene Softwareprodukte erworben. In die Erprobung geeigneter Methoden und Verfahren waren die Softwaresysteme Averbis Extraction Platform der Averbis GmbH, TopicFinder und iFinder der Intrafind Software GmbH, iSquare SmartSearch der iSquare GmbH und RapidMiner der Rapid-I GmbH einbezogen. Die Anbieter dieser Systeme haben die Tests bei der DNB ein Jahr lang aktiv begleitet. Die Auswahl der Erschließungssoftware wurde über ein europaweites Ausschreibungsverfahren getroffen. Mit Blick auf die spätere Integrierbarkeit in die Systemarchitektur der DNB spielten bei der Auswahlentscheidung auch technische Eigenschaften eine wichtige Rolle. Verlangt wurde ein hochgradig offener, modularer und transparenter Systemaufbau. Für den Titel- und Normdatenabgleich werden Funktionen des vorhandenen zentralen Bibliothekssystems benutzt.

Automatischer Titel- und Normdatenabgleich

Ein erster wichtiger Schritt zur Verringerung der eigenen intellektuellen Erschließungsaufwände ist eine weitgehende Nachnutzung schon vorhandener Metadaten bei der Erstellung von Titel- und Normdatensätzen. Strukturiert gelieferte Titeldaten der Netzpublikationen sollen auch dann übernommen werden, wenn diese nicht vollständig den Bestimmungen der derzeit angewandten Regelwerke entsprechen oder über deren Umfang hinausgehen. So werden beispielsweise alle Verfasser einer Publikation dauerhaft in den Titeldatensatz übernommen, und zwar auch dann, wenn es sich um eine größere Anzahl handelt. Außerdem sollen die Verknüpfungen zwischen den Titeldaten und den Normdateien ausgebaut werden, um Semantic Web-Anwendungen besser unterstützen und den Nutzern weitere Verknüpfungen für die Suche zur Verfügung stellen zu können.

Nachnutzung vorhandener Metadaten

In PETRUS wird zurzeit daran gearbeitet, Inhaltserschließungsdaten und Normdatenverknüpfungen aus bereits vorhandenen Datensätzen zu übernehmen, wenn parallele oder ähnliche Ausgaben als solche erkannt werden. Der Abgleich soll möglichst auch bei unterschiedlichen Schreibweisen oder Schreibfehlern funktionieren. Die Datensätze »echter« paralleler Ausgaben – beispielsweise Online- und Printausgaben desselben Werkes – werden direkt miteinander verknüpft. Keine »echten« Parallelausgaben im engeren Sinne sind Buchhandels- und Hochschulausgaben eines Werkes, unterschiedliche Auflagen eines Werkes oder Ausgaben eines Werkes in verschiedenen Verlagen. Vorhandene Übereinstimmungen können allerdings genutzt werden, um Verknüpfungen mit der PND und GKD oder Inhaltserschließungsdaten – beispielsweise die Notationen und Sachgruppen der Dewey-Dezimalklassifikation Deutsch (DDC) oder Schlagwörter – zwischen den Titeln auszutauschen.

Automatischer Titelabgleich paralleler Ausgaben

Im Zuge mehrerer Testreihen wurde untersucht, wie der Titelabgleich so gestaltet werden kann, dass möglichst keine falschen Erschließungsergebnisse produziert und dennoch ein möglichst großer Effekt erzielt werden kann. Entwicklung und Test der ersten Stufe sind nunmehr weitgehend abgeschlossen. Die Erkennung paralleler Ausgaben soll sowohl rückwirkend für die bereits in der bibliografischen Datenbank vorhandenen Netzpublikationen angewendet als auch als neues Modul in den laufenden Geschäftsprozess eingefügt werden. In weiteren Entwicklungsstufen soll die Erkennung von Ähnlichkeiten für den Zweck der Datenübernahme und Datenbereinigung – z. B. die Erkennung von Dubletten – weiter ausgebaut werden.

Auch zur Erstellung von Relationen zwischen Titeldaten und Normdaten werden Module benötigt, die Ähnlichkeiten erkennen. Beginnend mit den Personennamen in der PND wird beim Import der Netzpublikationen automatisch überprüft, ob die verzeichneten Personennamen bereits in der PND existieren. Wenn dies nicht der Fall ist, erzeugt das System automatisch neue Normdatensätze mit Relationen zu den Titeldaten. Ansonsten erfolgt eine Verknüpfung mit schon vorhandenen Personennamen in der PND. Erste Ansätze sind realisiert. In weiteren Schritten soll das Verfahren ausgebaut werden und auch individualisierte Datensätze mit in den Abgleich einbeziehen, also Personennamen in der PND, die bereits mit Angaben zum Lebenslauf der Person angereichert sind. Eine Individualisierung der automatisch generierten Normdatensätze soll nur dann ausgelöst werden, wenn ein Personenname mit mehreren Titeln verknüpft ist. Nach bisherigen Erkenntnissen sind zurzeit etwa 40 % der Personennamen in der PND lediglich mit einem einzigen Titel verknüpft.

Module zur Erstellung von Relationen Titeldaten/ Normdaten

Erschließung mit DDC-Sachgruppen

Zur Erschließung von Titeldaten setzt die DNB eine Systematik ein, die sich an der DDC orientiert und die aus etwa hundert Sachgruppen besteht.

Nutzung der DDC-Sachgruppen zur automatischen Kategorisierung

Die DDC-Sachgruppen dienen der thematischen Ordnung der gesammelten Publikationen unabhängig von der Sprache der Publikation. Mit dieser Systematik sollen maschinenlesbare deutsch- und englischsprachige Publikationen künftig automatisch kategorisiert – also in die DDC-Sachgruppen eingeordnet – werden.

In der Testphase wurden maschinelle Lernverfahren erprobt, die mit bereits erschlossenen Publikationen trainieren. Für Training und Test stand ein Korpus mit etwa 45.000 digitalen Volltexten zur Verfügung. Es handelt sich um Online-Hochschulschriften und einige andere monografische Netzpublikationen. Als weitere Testobjekte wurden zusätzlich die seit Jahresbeginn 2010 gesammelten und bisher noch nicht erschlossenen Netzpublikationen berücksichtigt. Die starke Vorherrschaft bestimmter Fächer, insbesondere der Medizin bei den Dissertationen, hat eine sehr ungleiche fachliche Verteilung zur Folge. Bei den Hochschulschriften konzentrieren sich etwa 90 % der Publikationen auf lediglich 20 Sachgruppen. Diese Unausgewogenheit wirkt sich ungünstig auf den Trainingsprozess aus und führt dazu, dass die spezifischen Erkennungsmerkmale selten vorkommender Sachgruppen vom System nur schwer identifiziert werden können.

Testkorpus mit 45.000 digitalen Volltexten

Ungleiche fachliche Verteilung erschwert das Training

Vor diesem Hintergrund wurden auch maschinenlesbare Textelemente aus der Kataloganreicherung in die Versuche mit einbezogen. Die DNB selbst scannt seit etwa drei Jahren Inhaltsverzeichnisse gedruckter Monografien. Sie übernimmt zudem gescannte Inhaltsverzeichnisse von Bibliotheksverbünden. Während der Tests konnte auf ein Kontingent von etwa 120.000 digitalisierten Inhaltsverzeichnissen einschließlich der für das Training benötigten Angaben zu den Sachgruppen zurückgegriffen werden. Die beobachtete verzerrte Verteilung ist tendenziell auch hier festzustellen, allerdings weniger stark ausgeprägt. Aufgrund von zu kleinen Trainingsmengen konnten bei den Online-Hochschulschriften letztlich nur 45 Sachgruppen in die Untersuchungen einbezogen werden, bei den gescannten Inhaltsverzeichnissen gedruckter Monografien waren es immerhin 81 Sachgruppen. Trainiert wurde teilweise mit mindestens 50, teilweise mit mindestens 70 Beispielobjekten pro Sachgruppe.

Testkorpus mit 120.000 digitalisierten Inhaltsverzeichnissen

Im Zuge der Experimente wurden die Ergebnisse der automatischen Sachgruppenvergabe mit den von Fachreferenten zugeordneten Sachgruppen verglichen, um verschiedene Maße zur Beurteilung der Qualität zu bestimmen. Die DNB will für den Geschäftsprozess der automatischen Erschließung von Netzpublikationen mindestens ein Niveau von 80 % korrekt kategorisierter Publikationen erreichen. Festgestellte Unterschiede zwischen intellektueller und automatischer Erschließung bedeuten nicht zwingend eine komplett falsche Kategorisierung, denn bei interdisziplinären Themenstellungen können auch mehrere DDC-Sachgruppen richtig sein. Die DNB vergibt in der Regel allerdings nur eine DDC-Hauptsachgruppe.

Automatische Sachgruppenvergabe im Test

Außerdem wurde im Rahmen der Tests eine mittlere Übereinstimmung der vergebenen Hauptsachgruppen von bis zu 75 % erreicht. Allerdings variieren die Ergebnisse von Sachgruppe zu Sachgruppe teilweise erheblich. Bei manchen Themen gelingt es den automatischen Verfahren sofort, eine nahezu korrekte Sachgruppenvergabe zu erreichen. Demgegenüber sind andere Themen deutlich schwerer voneinander abzugrenzen. Als nächster Schritt ist eine gezielte Erweiterung und Optimierung der Trainingskorpora geplant. Aufgrund der bisherigen Ergebnisse wird davon ausgegangen, dass die stetig wachsende Zahl gescannter Inhaltsverzeichnisse sehr gut für das Training der Klassifikatoren mit genutzt werden kann.

Trainingskorpora müssen erweitert und optimiert werden

Alle Softwaresysteme im Test haben zur Lösung der Aufgabenstellung letztlich auf Algorithmen aus dem Bereich der Support Vector Maschines (SVM) gesetzt. Beim Training der Klassifikatoren werden dabei hochkomplexe Teilungsfunktionen berechnet, die die Publikationen, die zu einer Sachgruppe gehören, von denjenigen trennen, die nicht dazugehören. Diese mathematischen Modelle werden anschließend für die Kategorisierung neuer Publikationen benutzt. Vorgeschaltet sind Sprachverarbeitungskomponenten, die die Sprache der Texte identifizieren, die Texte in Wörter zerlegen, eine linguistische Bearbeitung durchführen und die relevanten Informationseinheiten extrahieren. Mit linguistischen Methoden werden sprachliche Variationen wie beispielsweise Wortbeugung, Wortableitung und Wortzusammensetzung normalisiert, d. h. auf ihre Grundformen zurückgeführt.

Sprachverarbeitungskomponenten und mathematische Modelle zur Kategorisierung

In die geplanten Geschäftsprozesse soll die automatische Sachgruppenvergabe als Webservice eingebettet werden. Texte und Titeldaten werden über definierte Schnittstellen an die Erschließungssoftware übergeben. Als Ergebnis wird eine Liste der wahrscheinlichsten Sachgruppen mit Aussagen zur Validität zurückerwartet.

Automatische Beschlagwortung

Ziel der automatischen Beschlagwortung ist es, deutschsprachige Netzpublikationen mit dem kontrollierten Vokabular von SWD, PND und später GND zu erschließen. Zunächst werden die Sachschlagwörter sowie die geografischen und ethnografischen Schlagwörter der SWD in das Erschließungssystem eingebunden. Die SWD umfasst zurzeit annähernd 170.000 Sachschlagwörter und mehr als 200.000 geografische und ethnografische Schlagwörter. Für englischsprachige Veröffentlichungen soll künftig ein entsprechendes Vokabular wie z. B. die Library of Congress Subject Headings (LCSH) eingesetzt werden. Um Nutzern weitere Sucheinstiege zu bieten, soll zusätzlich eine freie Beschlagwortung sowohl für deutsch- als auch für englischsprachige Publikationen realisiert werden.

Automatische Beschlagwortung mithilfe von Normdateien

Die Modelle für die automatische Beschlagwortung können nicht wie bei der Sachgruppenvergabe trainiert werden. Wegen der zu großen Zahl möglicher Schlagworte existiert keine geeignete Trainingsgrundlage mit einer ausreichenden Menge maschinenlesbarer Lernbeispiele für jedes Schlagwort. Linguistische Verfahren, Methoden und Konzepte werden damit zur entscheidenden Grundlage für die maschinelle Bearbeitung der Texte. Begriffe, die das System im Zuge der linguistischen Analyse als sinntragend identifiziert, sollen anschließend auf ihre Vorzugsbenennungen in der SWD zurückgeführt und mit Angaben zur Validität in einer sortierten Liste ausgegeben werden. Letztlich soll auch das Beschlagwortungsmodul als Webservice in die Geschäftsprozesse eingebunden werden.

Einsatz linguistischer Verfahren zur Schlagwortvergabe

In der Testphase wurden mit zwei Softwaresystemen erste Versuche durchgeführt. Dafür wurden Sachschlagwörter mit ihren Synonymen und Relationen aus der SWD in die Beschlagwortungsmodule eingelesen. Die Auswertung der Ergebnisse konzentrierte sich auf Stichproben aus wenigen gezielt ausgewählten Sachgruppen. Dabei wurde jedes einzelne der automatisch vergebenen Schlagworte von Fachreferenten auf einer Vier-Punkte-Skala als sehr nützlich, nützlich, wenig nützlich oder falsch beurteilt, um einen ersten Maßstab für die Bewertung der Qualität zu erhalten. Bei diesem Vorgehen können systematische Fehler gut erkannt und nachfolgend behoben werden. Auf andere Formen der Auswertung wie beispielsweise Retrievaltests wurde bewusst verzichtet.

Bewertung der Qualität durch Fachreferenten

Bei den Tests wurde die Anzahl der Schlagworte pro Publikation teilweise fest vorgegeben, teilweise wurde sie über Wahrscheinlichkeitsmaße gesteuert. Ein gutes Erschließungsergebnis liegt dann vor, wenn das System möglichst viele der gewünschten und keine unsinnigen Schlagworte ausgibt. Möglicherweise lässt sich die Trennung der geeigneten von den nicht geeigneten Schlagworten künftig über Validitätsmaße steuern. Weitere Untersuchungen sollen zeigen, ob dies möglich ist. Stärker noch als bei der Sachgruppenvergabe gestaltet sich die Optimierung der automatischen Beschlagwortung als ein schrittweiser experimenteller Vorgang, der voraussichtlich noch einige Zeit in Anspruch nehmen wird.

Validitätsmaße als Instrument der Steuerung

Eine besondere Herausforderung stellt die Disambiguierung dar, also die richtige Einordnung mehrdeutiger Begriffe, die in der SWD durch entsprechende Zusätze am Schlagwort unterschieden werden. Diese sollen künftig in die automatische Beschlagwortung mit einbezogen werden, um Homonyme oder Polyseme richtig einordnen zu können. Zusätzlich sollen die Inhaltserschließungsdaten vorhandener Titel aus der Datenbank sowie die Beziehungen zwischen den verschiedenen angewendeten Erschließungssystematiken – Sachgruppen, DDC-Notationen und Schlagworte – für die Optimierung der Verfahren mit genutzt werden. So soll z. B. mit bereits erschlossenen Titeln trainiert werden, welche Schlagworte typischerweise in welchen Konstellationen auftreten. Auch die durchgängig vorhandenen Relationen zwischen den Schlagwörtern aus der SWD und den DDC-Sachgruppen können für die Bestimmung von Plausibilitäten mit herangezogen werden. Annähernd 120.000 der Sachschlagwörter sind zudem mit DDC-Notationen verknüpft. Auch diese Relationen sollen ausgewertet werden.

Schwierige Auflösung von Mehrdeutigkeiten

Ausblick

Die Untersuchungen zur Realisierung von Geschäftsprozessen für die automatische Erschließung von Netzpublikationen werden jetzt nur noch mit einem System weitergeführt. Zu Beginn dieses Jahres wurde die Averbis Extraction Platform beschafft, um damit die automatische Kategorisierung nach DDC-Sachgruppen und die automatische Beschlagwortung für Netzpublikationen zu realisieren.

Anmerkung

1 Schöning-Walter, Christa: PETRUS. In: Dialog mit Bibliotheken, 22 (2010) 1, S. 15 - 19.