Christa Schöning-Walter

Automatische Erschließung – Herausforderung und Chance

Bericht über den PETRUS-Workshop

Im März 2011 fand in der Deutschen Nationalbibliothek (DNB) im Rahmen des PETRUS-Projektes[1)] ein Workshop zum Thema »Automatische Erschließungsverfahren« statt, bei dem etwa dreißig Vertreter aus Bibliotheken, Dokumentationseinrichtungen, Informationszentren und Forschungsinstituten ihre Konzepte und praktischen Erfahrungen vorstellten und diskutierten.[2)]

Anlass für die Anwendung automatischer Erschließungsverfahren ist fast immer die Notwendigkeit, stetig wachsende Informationsmengen effizient verarbeiten zu müssen, häufig verbunden mit Einsparungsmaßnahmen und einer angespannten Personalsituation. Automatische Verfahren sollen genutzt werden, um den intellektuellen Erschließungsaufwand zu reduzieren und Geschäftsgänge zu unterstützen oder neu zu gestalten. Im Workshop wurden Beispiele präsentiert, die darauf zielen, bereits vorhandene Erschließungsdaten automatisch zu übernehmen, Metadatensätze ohne weitere Sacherschließungsdaten zumindest in fachliche Kategorien einzuordnen oder Schlagwörter als zusätzliche Sucheinstiege für die Datenbankrecherche automatisch zu vergeben. Die Mehrzahl der Workshop-Teilnehmer berichtete über Maßnahmen, die sich noch mehr oder weniger im Projektstatus befinden. Andere Einrichtungen wie beispielsweise die Pressedokumentation des Zweiten Deutschen Fernsehens (ZDF) in Mainz oder das Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) in Trier verfügen allerdings bereits über mehrjährige Praxiserfahrungen.

Automatische Erschließungsverfahren zur Bewältigung wachsender Informationsmengen

Vorgestellt wurden verschiedene Ansätze zur Klassifizierung bzw. Indexierung ausgewählter Bestände mit sowohl linguistischen als auch statistischen Methoden. Die Vertreterin des ZDF, einer der Vorreiter auf diesem Gebiet, berichtete ausführlich über die Erfahrungen bei der Einführung und beim Einsatz einer Kategorisierungssoftware für die semi-automatische Verschlagwortung von Zeitungs- und Zeitschriftenartikeln in der Pressedatenbank. Die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW) in Kiel und Hamburg arbeitet in Zusammenarbeit mit der Recommind GmbH ebenfalls auf die Einführung eines probabilistischen Verfahrens hin, um auf der Grundlage der Begriffe des Standard Thesaurus Wirtschaft (STW) eine Annotation wirtschaftswissenschaftlicher Texte mit automatisch generierten Schlagwörtern zu erreichen. Das GESIS – Leibniz-Institut für Sozialwissenschaften in Bonn untersucht eine entsprechende Lösung für die Verschlagwortung sozialwissenschaftlicher Publikationen und das Deutsche Institut für Internationale Pädagogische Forschung (DIPF) in Frankfurt testet verschiedene Systeme und Lösungsansätze hinsichtlich ihrer Eignung für die Erschließung erziehungswissenschaftlicher Literatur. Demgegenüber hat sich das ZPID bereits vor einigen Jahren für das überwiegend regelbasierte Indexierungssystem AUTINDEX entschieden und generiert damit Vorschläge für die intellektuelle Verschlagwortung von Psychologie-Information.

Anwendung linguistischer sowie statistischer Methoden

Andere Einrichtungen verfolgen im Grundsatz ähnliche Ziele. Die Technische Informationsbibliothek (TIB) in Hannover beispielsweise erprobt verschiedene Kategorisierungssysteme mit dem Ziel, den wachsenden Bestand an Metadatensätzen ohne Sacherschließungsdaten für die Fachsuche künftig automatisch nach sechs Fachgebieten zu klassifizieren. Die DNB will im Rahmen von PETRUS sowohl eine automatische Kategorisierung von Netzpublikationen nach DDC-Sachgruppen erreichen als auch eine automatische Annotation mit Schlagwörtern auf Basis der Normdateien, die zurzeit zur Gemeinsamen Normdatei (GND) zusammengeführt werden. Sie setzt zum einen auf ein statistisches Verfahren mit etwa hundert Klassen, zum anderen auf ein überwiegend linguistisches Verfahren mit einem Wörterbuch, das mehrere hunderttausend Sachschlagwörter, Geografika, Ethnografika, Personennamen etc. umfasst. Die Averbis

GmbH als Softwarepartner stellte die zugrundeliegenden Textanalyse-Verfahren im Workshop vor. In der Schweiz bereitet sich die Zentralbibliothek Zürich auf ein Pilotprojekt für die computerunterstützte Sacherschließung in drei Fachgebieten vor, das insbesondere auch die Mehrsprachigkeit des Landes von vornherein mit im Fokus hat. Auf diesem Gebiet arbeitet auch die Eurospider Information Technology AG, Zürich, die z. B. einen Rechtsthesaurus für mehrere Sprachen entwickelt hat.

Daneben wurden aber auch andere Technologien vorgestellt. Ein Vertreter der Universitätsbibliothek Mannheim berichtete über erfolgreiche Experimente mit Methoden des fallbasierten Schließens, angewendet auf Titeldaten verschiedener Bibliotheksverbünde, um Erschließungsdaten aus Ausgaben zu übernehmen, die als ähnlich erkannt werden. Damit können beispielsweise noch nicht oder unterschiedlich erschlossene Bestände einheitlich nach der Regensburger Verbundklassifikation (RVK) klassifiziert werden. Auch die Übernahme anderer Erschließungsdaten erscheint möglich. Außerdem wurde am Beispiel von SEMTINEL[3] über Forschungsarbeiten zur Analyse und Visualisierung von Thesauruskonzepten berichtet. Diese Software bietet interessante Techniken, um Strukturen sichtbar zu machen und so die Qualität eines Thesaurus zu überwachen. Der Vertreter des CONTENTUS-Projektes[4] gab einen Einblick in die technologischen Herausforderungen, die sich beim Umgang mit multimedialen Archiven stellen.

Vorstellung anderer Technologien

Viele der vorgestellten Erschließungssysteme setzen – zumindest teilweise – auf Methoden des maschinellen Lernens. Trainiert wird mit bereits erschlossenen Beispielen. Ganz entscheidende Faktoren für gute Ergebnisse bei der automatischen Erschließung sind das Vorhandensein ausreichender Trainingsdaten und die ausgewogene Auswahl repräsentativer Beispiele. Je mehr geeignete Trainingsobjekte dem System zur Verfügung stehen, desto erfolgreicher ist in der Regel die thematische Einordnung oder die Annotation von Schlagwörtern. In der Praxis ist dies typischerweise nicht für alle Kategorien gleichermaßen der Fall. Dann muss eventuell auch eine Anpassung der Taxonomie in Betracht gezogen werden. Es wurde berichtet, dass teilweise bessere Erschließungsergebnisse erzielt wurden, wenn Abstracts und nicht Volltexte analysiert werden, weil in der Zusammenfassung meistens das treffendste Vokabular verwendet wird. Die komplette Analyse langer Volltexte verursacht demgegenüber erfahrungsgemäß einen enormen maschinellen Aufwand ohne einen entsprechenden Nutzen. Deshalb werden in der Praxis üblicherweise nur Ausschnitte verarbeitet. Allerdings kann es bei heterogenem Material durchaus schwierig sein, diese Ausschnitte optimal festzulegen.

Maschinelles Lernen mit Trainingsdaten

Nicht immer stehen überhaupt maschinenlesbare Abstracts oder Volltexte für Training, Test und Produktion zur Verfügung. Manche Anwendungen, z. B. das schon genannte Vorhaben der TIB, setzen deshalb ausschließlich auf die Analyse von Metadaten. Die DNB nutzt demgegenüber elektronische Volltexte und zusätzlich digitalisierte Inhaltsverzeichnisse gedruckter Monografien, um Trainingskorpora für die automatische Sachgruppenvergabe aufzubauen. Das Trainingsmaterial sollte allerdings in seinen wesentlichen inhaltlichen Merkmalen mit den Publikationen übereinstimmen, die später produktiv verarbeitet werden. Schwierigkeiten bereiten auch bestimmte Literaturgattungen: sehr allgemeine Texte sind ebenso schwer einzuordnen wie fachlich sehr spezielle – einmalige – Publikationen. Texte mit vielen Formeln oder Sonderzeichen wiederum erfordern eine besondere Aufbereitung.

Qualitätsbewertung mit automatischen und intellektuellen Verfahren

Auch zur Beurteilung der Qualität der automatischen Erschließungsergebnisse im Testbetrieb werden – wo möglich – intellektuell vergebene Schlagwörter, Deskriptoren oder Klassen für einen Ergebnisvergleich herangezogen. Sind solche Metadaten vorhanden, dann ist teilweise ein maschineller Abgleich zur Ermittlung von Richtigkeit, Vollständigkeit und Nützlichkeit der automatischen Erschließungsergebnisse möglich. Der Abgleich stößt an Grenzen, wenn mehrere oder sogar verschiedene Zuordnungen richtig sind. Dennoch lassen sich auf diese Weise vielleicht Trends sichtbar machen. Vielfach führt jedoch kein Weg an einer intellektuellen Betrachtung mehr oder weniger großer Stichproben vorbei, um die Güte der Ergebnisse zu beurteilen. So unterschiedlich die Ziele und Rahmenbedingungen der einzelnen Einrichtungen sind, so individuell ist auch die geschilderte Vorgehensweise bei der Auswertung von Testreihen.

Aufwendige Vorarbeiten bis zur Inbetriebnahme zur Anpassung des verwendeten Vokabulars

Bis zufriedenstellende Ergebnisse erreicht werden, vergehen oft mehrere Jahre des Testens, Analysierens und Modifizierens. Vielfach müssen die Taxonomien, Wörterbücher oder Regeln in Stufen angepasst werden, um die Ergebnisse schrittweise zu verbessern. Berichtet wurde über teilweise sehr aufwendige Vorarbeiten bis zur Inbetriebnahme des automatischen Erschließungssystems, insbesondere in Bezug auf die Anpassung des verwendeten Vokabulars. Auch wenn die automatischen Verfahren bereits in die Geschäftsgänge integriert sind, muss beispielsweise der Thesaurus ständig weiter gepflegt werden.

Reduktion versus Erweiterung von Thesauri

So zwingt der statistische Ansatz erfahrungsgemäß zunächst zu einer Reduktion der verwendeten Klassen. Das ZDF beispielsweise hat seinen Thesaurus für die automatische Kategorisierung von zunächst etwa 3.500 auf schließlich etwa 2.000 Begriffe reduziert, GESIS zieht ein solches Vorgehen ebenfalls in Betracht. Auch in der Zentralbibliothek Zürich sollen zuerst gezielt die Sachschlagwörter für die automatische Erschließung selektiert werden. Im ZPID mit seinem linguistischen Verfahrensansatz wurden demgegenüber die im Thesaurus enthaltenen Begriffe durch über 18.000 Indikatoren zur präziseren Beschreibung der Deskriptoren und Synonyme ergänzt, um im Zuge der linguistischen Analyse eine bessere Disambiguierung zu erreichen. Eine besondere Herausforderung stellt auch die automatische Vergabe von sehr komplexen und spezifischen Schlagwörtern dar, wie z. B. das DIPF sie verwendet, oder die Vergabe von Deskriptoren aus einem polyhierarchischen Thesaurus wie dem STW. Die ZBW prüft deshalb, zur Analyse der Thesaurus-Strukturen das bereits erwähnte Evaluierungstool SEMTINEL einzusetzen.

Neben den manchmal sehr langwierigen Initialisierungsarbeiten wurden im Workshop auch die Grenzen der automatischen Erschließung deutlich angesprochen: den maschinellen Verfahren fehlt es häufig an Trennschärfe, eine bestimmte Fehlerquote muss daher in Kauf genommen werden. Auch ein gewisser Kontrollverlust muss akzeptiert werden, weil bei der Verarbeitung von Massendaten außer in Stichproben keine umfassende Prüfung der Ergebnisse möglich ist. Da die einzige Alternative aber vielfach nur darin besteht, dass gar keine Erschließung durchgeführt werden kann, stellt der Einsatz automatischer Kategorisierungs- oder Indexierungsverfahren auf jeden Fall einen Fortschritt dar. Die Systeme liefern i. d. R. zusammen mit den Erschließungsergebnissen auch Aussagen zur Wahrscheinlichkeit der richtigen Zuordnung. Daraus ergeben sich wiederum Optionen, die Fehler im Geschäftsgang durch angepasste Qualitätssicherungsmaßnahmen auf ein möglichst geringes Maß zu reduzieren.

Grenzen der automatischen Erschließung

Die beiden produktiven Anwendungen, über die im Workshop berichtet wurde, sind semi-automatische Verfahren zur Unterstützung konventioneller Geschäftsgänge. Erzeugt werden Vorschläge zur Beschleunigung, Vereinfachung und Vereinheitlichung der intellektuellen Erschließung. Sowohl das ZDF als auch das ZPID berichteten über eine gestiegene Effizienz und auch über eine wachsende Zufriedenheit der betroffenen Mitarbeiterinnen und Mitarbeiter nach anfänglicher Skepsis. Einige der künftigen Anwendungen, so wie sie beispielsweise bei der TIB oder der DNB geplant sind, streben von vornherein eine noch weitergehende Automatisierung an.

Ausblick

Der Workshop »Automatische Erschließungsverfahren« bei der DNB war die zweite Veranstaltung dieser Art nach einem Zusammentreffen beim DIPF etwa ein Jahr zuvor. Der Informations- und Erfahrungsaustausch in dieser Form stößt auf ein großes Interesse und soll nach Möglichkeit auch in den nächsten Jahren wiederholt werden.

Anmerkungen

1 <http://www.dnb.de/wir/projekte/petrus.htm>

2 Die Beiträge sind unter <http://www.dnb.de/wir/projekte/workshop_petrus.htm> online verfügbar.

3 <http://www.semtinel.org>

4 <http://www.d-nb.de/wir/projekte/contentus.htm>