Automatische Indexierung auf Basis von Titeln und Autoren-Keywords – ein Werkstattbericht

Martin Toepfer[1] und Andreas Oskar Kempf[2]

[1]*Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW),*
m.toepfer@zbw.eu
[2]*Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW),*
a.kempf@zbw.eu

Abstract

Automatische Verfahren sind für Bibliotheken essentiell, um die Erschliessung stetig wachsender Datenmengen zu stemmen. Die Deutsche Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft sammelt seit Längerem Erfahrungen im Bereich automatischer Indexierung und baut hier eigene Kompetenzen auf. Aufgrund rechtlicher Restriktionen werden unter anderem Ansätze untersucht, die ohne Volltextnutzung arbeiten. Dieser Beitrag gibt einen Einblick in ein laufendes Teilprojekt, das unter Verwendung von Titeln und Autoren[1]-Keywords auf eine Nachnormierung der inhaltsbeschreibenden Metadaten auf den Standard-Thesaurus Wirtschaft (STW) abzielt. Wir erläutern den Hintergrund der Arbeit, betrachten die Systemarchitektur und stellen erste vielversprechende Ergebnisse eines dokumentenorientierten Verfahrens vor.

Automatic systems are indispensable for libraries in order to make the rapidly growing number of publications accessible to their users. In the past the ZBW – German National Library of Economics – Leibniz Information Centre for Economics has gained practical experience in this field. Due to legal constraints it currently investigates methods that solely use author generated descriptive metadata. This article gives an insight into on-going developments and relates them to past activities. We report on a promising document-oriented approach, which uses author keywords and titles in combination to automatically assign subject headings from the STW Thesaurus for Economics to a document.

1 Einleitung

Bibliotheken und Infrastruktureinrichtungen sammeln bereits seit Längerem Erfahrungen im Bereich der automatisch gestützten Literaturerschliessung. Die Gründe liegen vor allem in steigenden Publikationszahlen und in veränderten Ausgangs- und Rahmenbedingungen der Inhaltserschliessung selbst. Rechtliche Aspekte müssen bei der Verarbeitung berücksichtigt werden. Insbesondere dürfen Systeme in der Praxis häufig nicht auf die zentrale Ressource, den Volltext, zugreifen.

[1]Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung männlicher und weiblicher Sprachformen verzichtet. Sämtliche Personenbezeichnungen gelten für beide Geschlechter.

Die Zunahme wissenschaftlicher Erkenntnisse (Bornmann und Mutz 2015), die Expansion der Forschungssysteme zahlreicher Länder (Alexander von Humboldt-Stiftung 2009) und der hohe Publikationsdruck, gemäss dem Diktum des „*Publish-or-Perish*“, haben in den vergangenen Jahren zu einem enormen Anstieg an Publikationen geführt. Klar erkennbar ist der Wandel von der gedruckten zur digitalen Veröffentlichung. Zusätzlich zeichnen sich die unterschiedlichen Wissenschaftsdisziplinen mit ihren fachspezifischen Publikationskulturen durch jeweils eigene Dynamiken im Publikationsverhalten aus. So zeigt sich etwa in den Wirtschaftswissenschaften eine klare Tendenz hin zur Veröffentlichung in wissenschaftlichen Journalen, was die Zahl der in Koautorschaft veröffentlichten Arbeiten geradezu „explodieren“ (Leininger 2009:68) liess. Die Publikationsformen Buch und Konferenzbeitrag hingegen erfahren einen relativen Bedeutungsverlust (Leininger 2009:68).

Daneben haben sich die Ausgangs- und Rahmenbedingungen der Inhaltserschliessung verändert. Digitale Veröffentlichungen verfügen neben dem Volltext vielfach bereits über inhaltsbeschreibende Metadaten in digitaler Form, die sich dank neuer Entwicklungen und Verfahren im Bereich des Text- und Data-Mining in bisher ungewohnter Art und Weise weiterverarbeiten und für das Retrieval nutzen lassen. So verfügt ein Grossteil der Zeitschriftenaufsätze etwa über Autoren-Keywords, die zum Teil in den Katalogisaten mitgeliefert werden und die urheberrechtlich nicht geschützt sind (Klimpel 2015). Der Anteil an inhaltsbeschreibenden Metadaten, die ausserhalb von Bibliotheken generiert wurden und nachgenutzt werden können, steigt somit. Zusätzlich bilden einzelne Dokumentenkollektionen immer mehr lediglich einen Teilbestand von Portalen und Discovery Systemen, wodurch die Heterogenität der Erschliessungsdaten weiter zunimmt.

Vor diesem Hintergrund erscheinen die Mehrwerte klassischer bibliothekarischer Erschliessungsinstrumente, wie kontrollierte Vokabulare, ungebrochen. Thesauri und Klassifikationssysteme normieren das Erschliessungsvokabular. Sie helfen, die Benennungsvielfalt (Synonymie) und Mehrdeutigkeit (Polysemie) von Sprache zu kontrollieren und Inhalte strukturiert zu erfassen. Sie steigern die Indexierungskonsistenz und bieten im Falle eines umfangreichen Zugangsvokabulars insbesondere dann einen Mehrwert bei der Suche, wenn Suchbegriffe nicht direkt in Titel, Abstract oder Volltext enthalten sind. Die Anreicherung von Indexaten mit einem kontrollierten Vokabular erleichtert das Retrieval, ermöglicht einen schnellen Zugang und beschleunigt die Relevanzentscheidung deutlich (Bertram 2005). Im Fall von unterschiedlichen Quellen inhaltsbeschreibender Metadaten erwächst ein besonderer Bedarf, die heterogenen Inhaltsdaten zu harmonisieren.

Mit dem Einsatz maschineller Verfahren werden häufig verschiedene Teilziele verfolgt. Zum einen geht es darum, Werke zu erschliessen, die bisher aus Ressourcengründen nicht erschlossen werden konnten. Vorteil maschineller Verfahren ist ihre Anwendung auf grosse Datenmengen. Zum anderen geht es um die Nachnutzung von Fremddaten unterschiedlicher Art, um die heterogenen inhaltsbeschreibenden Metadaten einander anzugleichen. Während für die Nachnutzung fremder bibliothekarischer Normdaten Crosskonkordanzen aufgebaut werden, gilt es, die nicht-bibliothekarisch erzeugten Daten auf das eigene kontrollierte Vokabular abzubilden. Diesen Prozess, Katalogeinträge, in denen Autoren-Keywords vorhanden sind, in Katalogeinträge mit passenden Begriffen (Deskriptoren) aus einem kontrollierten Vokabular umzuwandeln, verstehen wir als Nachnormierung. Bei dieser Transformation können Probleme wie Mehrdeutigkeiten auftreten, bei denen möglichst genaue Beschreibungen nur mit entsprechendem Kontext und komplementären Quellen möglich sind.

Im Folgenden erläutern wir zunächst den Hintergrund der aktuellen Arbeit. Wir beziehen uns auf Erfahrungen mit maschinellen Verfahren allgemein und an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW) – Leibniz-Informationszentrum Wirtschaft im Speziellen. Im Anschluss geben wir einen konkreten Einblick in ein laufendes Teilprojekt, bei dem die Systemarchitektur der Automatik gegenüber früheren Arbeiten Titel und Autoren-Keywords gemeinsam verwendet, um eine Nachnormierung auf den Standard-Thesaurus Wirtschaft (STW) zu erzielen. Im Gegenssatz zu einer statischen Verknüpfung im Sinne einer Crosskonkordanz bzw. Vokabularabbildung ist das jetzt verfolgte Vorgehen dokumentenorientiert und damit in der Lage, kontextbezogene Zuordnungen vorzunehmen. Der Artikel stellt neben der Systemarchitektur auch erste experimentelle Ergebnisse vor, die im Vergleich zu titelbasierten Vorhersagen bereits deutliche Verbesserungen aufzeigen.

2 Hintergrund

Die ZBW erfährt als weltweit grösste Infrastruktureinrichtung für wirtschaftswissenschaftliche Literatur mit einem klar überregionalen Auftrag in besonderer Weise die Auswirkungen des Publikationsanstiegs. Hochwertig aufbereitete Metadaten einschliesslich einer normierten inhaltlichen Beschreibung, die einen einfachen und schnellen Zugang zum Bestand ermöglicht, stellen für die ZBW einen entscheidenden Aspekt im Wettbewerb mit anderen Informationsdienstleistern dar (Groß und Faden 2010:1120). Der Zuwachs an Publikationen führt zu einem immer geringeren Sacherschliessungsgrad für die neueste wirtschaftswissenschaftliche Literatur (Wortmann, Groß und Bahls 2014). Der Anteil des eigenen kontrollierten Vokabulars, des STW, in der eigenen Datenbasis ist somit rückläufig. Gleichzeitig verfügt ein Grossteil der inhaltlich ausgewerteten Zeitschriften über eine freie Verschlagwortung, die von den Autoren selbst vorgenommen wurde.

Vor diesem Hintergrund sammelt die ZBW bereits seit Längerem Erfahrungen mit dem Einsatz maschineller Erschliessungsverfahren. Zwei verschiedene Phasen lassen sich unterscheiden. In der ersten Phase wurde ein software-getriebener Ansatz verfolgt. Den Auftakt bildete ein DFG-Projekt in den Jahren 2002-2004. In einer Kooperation mit dem Institut für angewandte Informationsforschung der Universität des Saarlandes, bei der die dort entwickelte semiautomatische Indexierungskomponente AUTINDEX[2] (Haller, Ripplinger und Maas 2000) zum Einsatz kam, wurden zentrale Anforderungen und auftretende Problemlagen eruiert. Eine Neuauflage erfuhr die Verwendung maschineller Verfahren mit dem Beschluss der ZBW im Jahr 2008, erneut ein Projekt zur Vorbereitung der Einführung einer automatischen Indexierung zu starten. Zwischen 2009 und 2011 ging die ZBW eine Zusammenarbeit mit der Firma Recommind ein, bei der der Einsatz der Erschliessungssoftware MindServer getestet wurde und auch Keywords gemeinsam mit Abstracts in einer dokumentenorientierten Weise eingesetzt wurden (Faden und Groß 2011).

Nach dieser ersten Lernphase wurde die weitere Beschäftigung mit automatischen Erschliessungsverfahren konzeptionell neu ausgerichtet. Aus den Erfahrungen mit extern entwickelten Software-Lösungen wird nun als Ziel verfolgt, eine eigene maschinelle Erschliessungsinfrastruktur aufzubauen. Diese Neuausrichtung ist mit dem Ausbau der ZBW zu einer Forschungs-

[2]Dieses Projekt wurde von dem damals noch existierenden Hamburgischen Weltwirtschaftsarchiv (HWWA) und der ZBW (Kiel) gemeinsam mit dem Saarbrücker Institut für Angewandte Informationsforschung (IAI) an der Universität des Saarlandes als Projektpartner durchgeführt.

bibliothek verbunden. Durch den Aufbau des Programmbereichs Medieninformatik werden eigene Kompetenzen im Bereich Text- und Data-Mining aufgebaut, die als Wissenstransfer in forschungsbasierte anwendungsorientierte Bibliotheksservices einfliessen. Daneben erfolgte mit Verabschiedung eines neuen Erschliessungskonzepts (Kempf und Rebholz 2016) eine Neuausrichtung der Sacherschliessung, die den veränderten Rahmenbedingungen der Erschliessung Rechnung trägt. Konzeptionell werden die Aktivitäten entsprechend eines Mehr-Ebenen-Ansatzes stärker als zuvor an den jeweiligen Publikationseigenschaften bezüglich Erscheinungsform, Verfügbarkeit nachnutzbarer Metadaten und Volltext-Verfügbarkeit ausgerichtet. Ziele sind unter anderem die Workflow-Unterstützung für Indexierer und die Anwendung eines maschinellen Verfahrens im Produktivbetrieb bei Teilbeständen.

Untersuchungen zum Vorkommen von STW-Begriffen in Autoren-Keywords und zur Nachnormierung durch Zuordnung im Sinne einer Crosskonkordanz wurden bereits 2014 und 2015 an der ZBW betrieben (Wortmann, Groß und Bahls 2014; sowie ein interner Praktikumsbericht aus dem Jahr 2015). Die Ergebnisse legten die Notwendigkeit des Einbezugs von Titeln nahe. Fremddaten wurden darüber hinaus auch als Quelle neuer Begriffe und Synonyme für den STW in Betracht gezogen (Rebholz und Bahls 2015).

Seit 2014 wurde zudem intensiv mit der Arbeitsgruppe *Knowledge Discovery* der Christian-Albrechts-Universität zu Kiel zusammengearbeitet und aktiver Forschungstransfer betrieben. Im Rahmen einer Masterarbeit (Große-Bölting 2015) wurden mehrere Verfahren und Konfigurationen eines Frameworks zur automatischen Indexierung von Volltexten verglichen. Insbesondere wurden auch hierarchie- und graphbasierte Aktivierungsfunktionen untersucht. Entwicklungen durch weitere studentische Tätigkeiten wurden unter dem Namen „quadflor“[3] (Schelten u. a. 2016) verfügbar gemacht. „Quadflor“ enthält eine methodische Erweiterung[4] und erzielt damit gute Ergebnisse bereits mit Titeln. Ebenfalls im Rahmen einer studentischen Arbeit wurden für einen Vergleich dazu die auf Ergebnissen der University of Waikato beruhende Open-Source Software „Maui“[5] (Medelyan und Witten 2006) auf Daten der ZBW evaluiert und Parameteroptimierungen durchgeführt.

Neben innovativen automatischen Indexierungsverfahren nehmen aktuell Qualitäts- und Indexierungskonsistenzmessungen einen hohen Stellenwert ein, um homogene Indexierungsergebnisse und eine hohe Metadatenqualität sicherzustellen.

3 Automatisches Indexierungssystem

In diesem Abschnitt beschreiben wir eine Systemkomponente zur automatischen Erschliessung, die derzeit an der ZBW untersucht und getestet wird. Wir gehen auf allgemeine Aspekte sowie auf Titel und Autoren-Keywords ein. Zuvor betrachten wir ein Beispiel, das illustriert, welcher Prozess durch die Automatik realisiert werden soll.

Abbildung 1 zeigt schematisch einen Fall aus der Praxis. Von den Verfassern des Werks liegen die inhaltsbeschreibenden Felder Titel („*International Technology* ...“) und Keywords

[3] https://github.com/quadflor/Quadflor

[4] „The most notable contribution is a stacked classifier called LRDT, which consists of stochastic gradient descent (optimizing logistic regression) and decision trees.“ https://github.com/quadflor/Quadflor

[5] https://github.com/zelandiya/maui

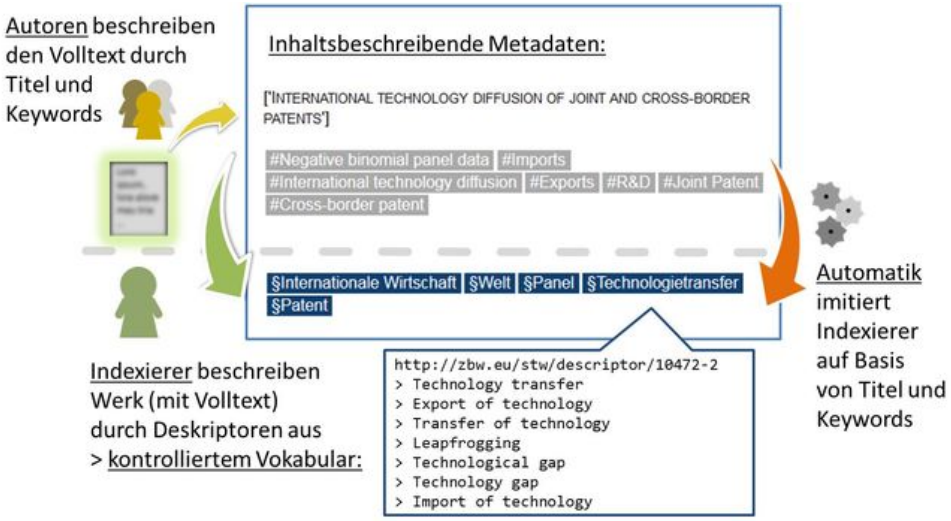

Abbildung 1: Schematische Abbildung einer dokumentenorientierten Transformation von Titel und Autoren-Keywords in Einträge eines kontrollierten Vokabulars

(„*#Negative binomial panel data*“, ...) vor, die auf Einträge des STW abgebildet werden („§Internationale Wirtschaft“, ...). Während die Indexierer dafür auch auf den Volltext zugreifen, benutzt die Automatik lediglich inhaltsbeschreibende Metadaten. Über die Zuordnung zu dem kontrollierten Vokabular und die dort hinterlegten Synonyme („§Technologietransfer“, Deskriptor 10472-2, Synonyme: „Export of technology“, ...) kann das Retrieval unterstützt werden. Zudem können Inhalte unterschiedlicher Datenbanken auf diese Weise verknüpft und strukturiert analysiert werden.

Das dargestellte Vorgehen, die Metadaten für die Indexierung zu nutzen, ist dokumentenorientiert. Das heisst, Autoren-Keywords werden nicht im Sinne einer Crosskonkordanz immer auf dieselben Deskriptoren abgebildet. Die Indexierung geschieht pro Metadatensatz eines Titels und kann dadurch den Kontext der Verwendung eines Autoren-Keywords berücksichtigen. „*#Virus*“ als Autoren-Schlagwort könnte somit beispielsweise auf medizinische oder informationstechnische Deskriptoren abgebildet werden, in Abhängigkeit der im Umfeld des Terms angegebenen Wörter. Eine starre Zuordnung könnte in diesem Fall schädlich sein.

3.1 Allgemeiner Überblick

Die Verarbeitung automatischer Indexierungssysteme beinhaltet in der Regel viele modular aufgebaute Schritte, die häufig auch Teil allgemeiner Verarbeitungssysteme natürlicher Sprache (Jurafsky und Martin 2009) sind. Automatische Indexierung ist insbesondere ähnlich zu Verfahren, die Dokumente thematisch kategorisieren bzw. gemäss einer Taxonomie klassifizieren. Pro Dokument können bei der Indexierung jedoch mehrere Deskriptoren relevant sein; man spricht von „*multi-label classification*“ (MLC). Die grundlegende Besonderheit gegenüber der Einordnung in genau eine von mehreren möglichen Klassen ist, dass bei MLC Abhängigkeiten zwischen den Verteilungen der Klassen bzw. Deskriptoren bestehen können. Bei einem „*one-vs-rest*“ Ansatz werden diese Abhängigkeiten wiederum ignoriert, da für jede Klasse ein eigener Binärklassifikator gebildet wird. Diese Klassifikatoren unterscheiden jeweils eine Klasse (*one*) von allen anderen (*rest*).

Für den Erfolg des Gesamtsystems sind viele Faktoren massgeblich. Charakteristika sind unter anderem a) die genutzten Eingabeinformationen zu den Dokumenten (z.B. Titel, Keywords, Volltext, etc.), b) die Transformation der Eingabe in Merkmale (engl.: *features*, z.B. Häufigkeit von n-Grammen), c) die Methoden, die zur Vorhersage und zum Schätzen von Parametern benutzt werden, und ihre Konfiguration sowie d) die Verfügbarkeit und Qualität von Wissensressourcen, wie beispielsweise Wörterbücher, Synonymbeziehungen, Crosskonkordanzen, und nicht zuletzt auch Beispieldokumente.

3.1.1 Genutzte Eingabeinformationen zu den Dokumenten

Orientiert man sich für den Entwurf der Automatik am manuellen Indexierungsprozess, ist die Struktur der Eingabeinformationen relevant. Indexierer benutzen, um die relevanten Deskriptoren zu bestimmen, spezielle Strategien, die Metadaten sowie strukturelle Aspekte der Dokumente ausnutzen. Sie fokussieren auf Bereiche der Dokumente, die die wichtigsten Aspekte des Inhalts kurz und prägnant beschreiben. Dazu gehören Titel und Autoren-Keywords, aber auch Abstract, Einführung und Zusammenfassung.

Automatische Indexierungssysteme setzen oft längere Textabschnitte wie Abstracts oder Volltexte ein (Haller, Ripplinger und Maas 2000; Medelyan und Witten 2006; Faden und Groß 2011), manche Systeme funktionieren jedoch auch alleine auf Basis der Dokumententitel (Ferber 2005; Schelten u. a. 2016).

3.1.2 Transformation der Eingabe in Merkmale

Die Eingabeinformationen, insbesondere Textfelder, werden anschliessend in eine Merkmalsrepräsentation umgewandelt, die die Eigenschaften des Dokuments in einem numerischen Vektor darstellt, oft ausschliesslich mit binären Werten. Die textuellen Felder werden dazu in sogenannte Tokens unterteilt, im einfachsten Fall durch Trennen an Leerzeichen. Anschliessend können Folgen von Tokens zu n-Grammen zusammengefasst werden (*n-gram features*), oder in Wörterbüchern nachgeschlagen werden (*concept features*). Ihre Bedeutung für die Inhaltsbeschreibung kann durch statistische Berechnungen gewichtet werden, beispielsweise durch Masse, die die Häufigkeit des Vorkommens im Dokument zur Häufigkeit in mehreren bzw. allen verfügbaren Dokumenten in Relation setzen (z.B. *term-frequency inverse-document-frequency*, tf-idf).

3.1.3 Methoden zur Vorhersage und zum Schätzen von Parametern und ihre Konfiguration

Unter anderem wegen der grossen Anzahl an unterschiedlichen Termen und möglichen Deskriptoren werden die Zuordnungen zwischen den Eigenschaften der Dokumente und den Konzepten über datengetriebene Ansätze umgesetzt. Werden die Parameter des Modells nur mit Hilfe eines Trainingsdatensatzes, das heisst Einträgen für die „korrekte“ Deskriptoren[6] gegeben sind, geschätzt, spricht man von überwachten Verfahren. Werden zusätzlich Instanzen hinzugezogen, die nicht verschlagwortet sind, spricht man von halb-überwachtem Lernen. Einen guten Überblick zu einzelnen Methoden liefert beispielsweise die Dokumentation von scikit-learn[7].

[6] In der Regel gibt es mehrere unterschiedliche, jeweils für sich plausible Möglichkeiten, den Inhalt eines Dokuments mit kontrolliertem Vokabular zu beschreiben.

[7] http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

3.1.4 Verfügbarkeit und Qualität von Wissensressourcen

Massgeblichen Einfluss auf die Ergebnisse hat die Auswahl der Wissensressourcen, die in das System einbezogen werden. Wörterbücher und Synonymbeziehungen können das System dabei unterstützen, auch zuvor nicht in Trainingsdaten beobachtbare Zusammenhänge zwischen Termen und Konzepten einzubeziehen. Strukturiert vorliegende Beziehungen zwischen Begriffen, beispielsweise Hierarchien, können in Automatiken genutzt werden. Zentral für den Erfolg überwachter Lernverfahren sind schliesslich die Beispieldokumente (Trainingsdaten), die zur Optimierung von Parametern verwendet werden. Sie sollten die Grundgesamtheit, auf der das Verfahren eingesetzt werden soll, angemessen repräsentieren und für jede Klasse, für die Parameter geschätzt werden, Musterbeispiele bereitstellen. Zudem sollte man sich für die Beurteilung der Leistung des Systems darüber bewusst sein, wie unscharf die Kategorien in den Trainingsdaten repräsentiert sind. Überlappen sich Begriffe sehr stark oder gibt es keine aussagekräftigen Merkmale bzw. Merkmalskombinationen, die die Begriffe voneinander trennen, führt das zu inkonsistenten Vorhersagen, die als Fehler gemessen werden.

3.2 Aktuelles System

Die ZBW untersucht in Zusammenarbeit mit der Arbeitsgruppe *Knowledge Discovery* der Christian-Albrechts-Universität zu Kiel unterschiedliche Systemkonfigurationen. Aufgrund positiver Erfahrungen in bereits abgeschlossenen Untersuchungen (Schelten u. a. 2016) wird zunächst nach dem „*one-vs-rest*“ Ansatz verfahren. Die Aufgabe, einem Dokument eine Menge von Deskriptoren zuzuweisen, wird transformiert in viele Teilaufgaben. Pro Deskriptor des STW wird entschieden, ob der Begriff relevant ist oder nicht, wofür einzelne Klassifikatoren mit vielen Parametern benutzt werden. Als Klassifikationsverfahren wird logistische Regression eingesetzt (Schelten u. a. 2016), dabei werden die Parameter durch ein Gradientenabstiegsverfahren (*stochastic gradient descent*, kurz: SGD) optimiert.

Die automatisch generierten inhaltsbeschreibenden Einträge stammen aus dem von der ZBW herausgegebenen Standard-Thesaurus Wirtschaft[8]. Dieser enthält über 6.000 Deskriptoren in Deutsch und Englisch und in der englischen Version zusätzlich über 6.500 Synonymverweise. Aufgeteilt auf sieben Subthesauri deckt er alle ökonomischen Themenfelder und wichtigen benachbarten Sachgebiete inklusive Allgemeinwörtern und Geografika ab. Die Begriffsansetzungen sind über Beziehungen miteinander verknüpft. Zum einen sind symmetrische Verwandtschaftsbeziehungen (*related*) strukturiert erfasst, zum anderen sind die Begriffe als Ober- und Unterbegriffe innerhalb hierarchischer Beziehungen (*broader/narrower*) angeordnet. Eine besondere Herausforderung besteht darin, dass der Thesaurus polyhierarchisch aufgebaut ist. Ein Begriff kann zugleich mehreren Oberbegriffen zugeordnet sein.

Als Software-Bibliotheken verwenden wir bei den Experimenten das auf scikit-learn[9] basierende „quadflor“ (Schelten u. a. 2016) und eine alternative Anbindung von scikit-learn, bei der andere Parameter gesetzt werden.

[8] http://zbw.eu/stw/version/latest/about.de.html
[9] http://scikit-learn.org

3.3 Titel

Merkmale aus Titeln sind ein elementarer Bestandteil des Systems. Titel umfassen nur wenige Phrasen, die von den Autoren passend zum Inhalt des Artikels gewählt werden. Für automatische Verfahren stellen diese Eigenschaften sowohl Chancen als auch Risiken dar.

Auf der einen Seite unterstreichen Autoren durch die Erwähnung eines Begriffs im Titel die Relevanz des Konzepts für die Beschreibung des gesamten Werks. Die Automatik muss diesen Aspekt nicht in dem Ausmass selbst bestimmen, wie es bei der Verarbeitung von Volltexten der Fall ist. Zudem sind Grammatik und Vokabular im Gegensatz zu Alltags- beziehungsweise Fachsprache vereinfacht.

Auf der anderen Seite gibt es im Gegensatz zu Volltexten nicht mehrere Umschreibungen und Synonyme, die den gleichen Begriff eines Themas adressieren. Die Wahrscheinlichkeit, dass das automatische Verfahren mindestens eine der benutzten Bezeichnungen erkennt, wird somit verringert. Ausserdem stellen Titel durch die geringe Länge nur wenig Kontext bereit, um mehrdeutige Terme auf eine Wortbedeutung abzubilden, oder zu entscheiden, ob sich für einen Artikel besser ein Oberbegriff oder mehrere spezifischere Begriffe eignen. Schliesslich können manche der für die inhaltliche Erschliessung relevanten Begriffe im Titel unerwähnt bleiben. Titel sind nicht der Gegenstand, der erschlossen wird, sondern nur eine Beschreibung, die von den Autoren gewählt wurde, unter anderem um Aufmerksamkeit beim Leser zu erzeugen. Die gewählte Beschreibung könnte demnach sogar irreführend sein, wenn Titel und Volltext nicht zueinander passen.

3.4 Autoren-Keywords

Für Autoren-Keywords gelten prinzipiell auch die für Titel genannten Überlegungen. Sie verwenden nicht exakt das intendierte kontrollierte Zielvokabular, sondern auch eigene idiosynkratische Phrasen. Autoren-Keywords können den Inhalt ausführlicher beschreiben als der Titel und sollen von den Verfassern so gewählt sein, dass sie die relevanten Begriffe darstellen. Im Gegensatz zu Folksonomies, also sozialen Verschlagwortungsumgebungen, bei denen Benutzer persönliche Keywords (Tags) vergeben, ist die Intention von Autoren eine nach aussen gerichtete Beschreibung aller relevanten Inhalte des Objekts, die versucht, etabliertes Vokabular zu verwenden, um die adressierte Zielgruppe zu erreichen. Autoren-Keywords können aus diesem Grund auch – ähnlich zu irreführenden Titeln – potenziell schadhaft sein, wenn sie nicht die Kernaspekte des Volltexts widerspiegeln. Ursache dafür könnte beispielsweise sein, dass Autoren weniger relevante Nebenaspekte des Artikels verschlagworten; entweder unbewusst, oder bewusst, um mehr Sichtbarkeit zu erlangen.

Autoren-Keywords können aber auch neue relevante Inhaltsaspekte enthalten, wie beispielsweise „#*Negative binomial panel data*“ im Beispiel in Abbildung 1. Mit einer ausführlichen Beschreibung der verwendeten Vorverarbeitungsmethoden haben Haustein und Peters (2012) beispielsweise professionell erstellte Indexate (kontrolliertes Vokabular), Tag-Zuweisungen von Lesern (Folksonomy), inhaltsbeschreibende Felder von Autoren (Keywords, Titel, Abstract) und automatisch erstellte Keywords auf komplementäre Inhaltsaspekte hin untersucht. Nach Bereinigung und Normalisierung der Phrasen und Terme wurde auf der Ebene von Dokumenten (*docsonomy*) gezeigt, dass unterschiedliche Perspektiven des Inhalts durch die verschiedenen Beschreibungen abgedeckt werden.

Ein Ansatz, die Inhalte, die in Autoren-Keywords Erwähnung finden, zu erschliessen, wäre, eine Crosskonkordanz zwischen dem Vokabular der Autoren-Keywords und dem kontrollierten Vokabular zu erstellen. Dieses Vorgehen ähnelt Ontologie-Abbildungen, die jedoch auch Strukturähnlichkeiten nutzen können, und erschiene beim Einsatz effizient und gut überprüfbar. Auf der anderen Seite wäre diese Herangehensweise nicht in der Lage, unscharfe Formulierungen aufzulösen. Einige passende Begriffe ergeben sich zudem nur aus der Kombination mehrerer Schlagwörter, zum Teil sind sogar einzelne Wörter oder Wortbestandteile relevant. Aus diesen Gründen untersuchen wir eine andere, dokumentenorientierte Vorgehensweise, die den Kontext des zu indexierenden Objekts berücksichtigt.

3.5 Titel+Keywords

Um Autoren-Keywords und Titel gemeinsam für die Vorhersage von Deskriptoren zu nutzen, kann man vielfältige Strategien verfolgen und unterschiedliche Methodiken anwenden. Beispielsweise kann man auf direkte Weise vorgehen und, wie in Abbildung 2 veranschaulicht, eine Feldkonkatenation oder eine Merkmalskonkatenation durchführen. Bei der Feldkonkatenation werden Titel und Autoren-Keywords zu einem gemeinsamen textuellen Feld vereinigt, das anschliessend in einen Merkmalsvektor übertragen wird, der nicht zwischen Termen aus dem Titel und Termen aus den Keywords unterscheidet. Der Term „patent“ taucht in dem Beispiel sowohl in Titel und Autoren-Keywords insgesamt dreimal auf. Beim Feldkonkatenationsverfahren beträgt das entsprechende Häufigkeitsmerkmal demnach den Wert 3. Die Merkmalskonkatenation repräsentiert die Häufigkeiten getrennt: 1x Titel, 2x Autoren-Keywords. Dadurch entstehen insgesamt mehr Parameter, die die Zusammenhänge zwischen Merkmalen und Deskriptoren darstellen und für deren Schätzung Daten zur Verfügung gestellt werden müssen. Die Merkmalswerte sind zudem pro Dokument weniger stark ausgeprägt, wie im Beispiel beim n-Gramm „*technology diffusion*“ zu beobachten ist (2 bei Feldkonkatenation, jeweils 1 bei Merkmalskonkatenation). In diesem Beitrag benutzen wir den Feldkonkatenationsansatz, da für mehrere Tausend Deskriptoren Parameter geschätzt werden müssen.

4 Experiment

4.1 Ziele

Mit den folgenden Untersuchungen möchten wir abschätzen, ob sich bereits mit dem einfachen Verfahren, Merkmale aus Titeln und Keywords gemeinsam zu nutzen, bessere Ergebnisse erzielen lassen als rein titelbasiert.

4.2 Aufbau

Grundlage der Experimente war eine Abfrage von englischsprachigen Dokumenten aus EconBiz, für die neben Titel und Autoren-Keywords auch eine STW-Verschlagwortung vorlag. Bei mehrfach auftretenden Titeln wurde nur der erste verwendet. Dies führte zu 20.227 Dokumenten. Insgesamt gab es mehr als 150.000 Dokumente mit Autoren-Keywords. Die Menge der zu einem Dokument manuell vergebenen Deskriptoren aus der Datenbank bezeichnen wir als Referenzindexat. Ein Referenzindexat enthielt im Mittel 5,85 (+/- 1,84) Deskriptoren. Knapp über 90% der Dokumente hatten mindestens vier und maximal zehn Deskriptoren. Insgesamt

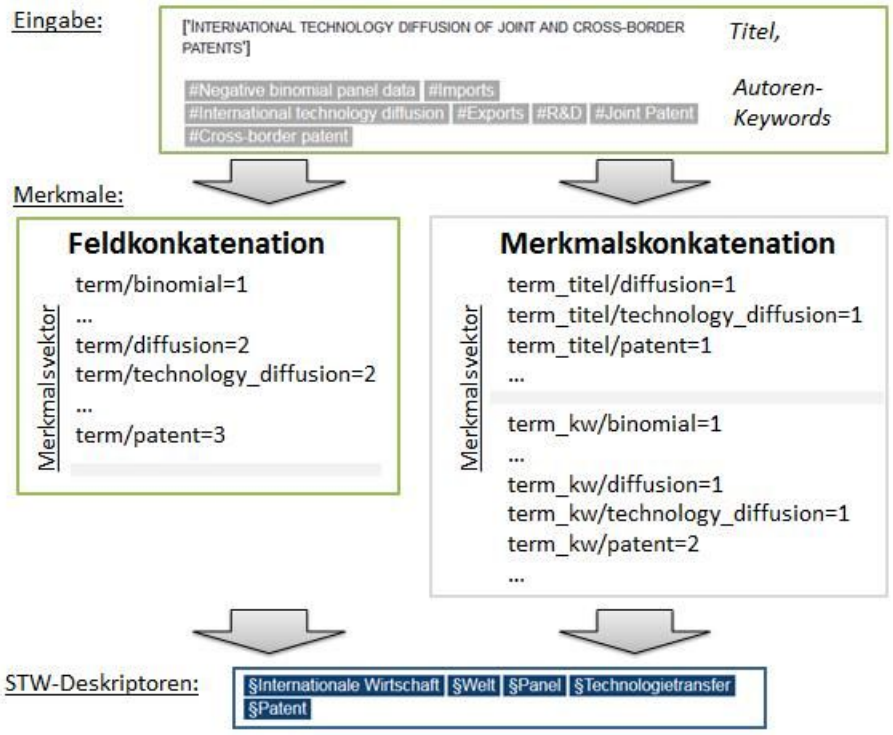

Abbildung 2: Unterschiedliche Möglichkeiten, Titel und Autoren-Keywords in Merkmalsvektoren zu überführen

werden 3.998 der über 6.000 Begriffe des STW verwendet. Der häufigste Deskriptor kam 4.678-mal vor. Über 2.000 Deskriptoren kamen weniger als zehnmal vor, wobei sie mit ca. 7% zur Gesamtanzahl an Deskriptorzuweisungen innerhalb des Dokumentenkorpus beitrugen.

Für die Bewertung der Verfahren benutzen wir etablierte Metriken (*precision*, *recall*, f1) und Verfahren (*k-fold cross validation*), wie sie häufig von Softwarebibliotheken bereitgestellt werden.[10] Die *precision* gibt prozentual an, wie oft die von der Automatik vorgeschlagenen Deskriptoren richtig waren. Der *recall* misst, wie viele der Deskriptoren aus den Referenzindexaten auch von der Automatik vorgeschlagen wurden. Das f1-Mass wird durch das harmonische Mittel der beiden Werte berechnet. Um einen Gesamtwert für mehrere Testdokumente und alle Deskriptoren zu erhalten, werden in diesem Beitrag die Masse zunächst pro Dokument berechnet und anschliessend gemittelt (*sample-based average*). Pro getesteter Konfiguration wurde die Gesamtmenge an Dokumenten in zehn Teile partitioniert. Jeder dieser Teile wurde anschliessend einmal als Testmenge verwendet und auf den restlichen neun Teilen trainiert. Schliesslich wurde jeweils über die zehn Durchläufe gemittelt.

Zunächst wurde „quadflor“ mit einer Konfiguration angewendet, für die optimale Ergebnisse berichtet wurden. Diese setzt einen Stacked-Classifier (Parameter „-f sgddt“, L2 penalty „-P l2“) ein und nutzt binäre Term- und Konzeptmerkmale (Parameter „-ctb“). Zusätzlich wurde eine einfachere Konfiguration (*basic*) benutzt, die weder einen Stacked-Classifier verwendet (Parameter „-f sgd“, *elastic net penalty* „-P elasticnet“) noch Konzeptmerkmale (Parameter „-tb“). Alle Konfigurationen verwenden den Alpha-Wert 1e-7.

[10] http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

In einem weiteren Versuch (Setting-B) wurde scikit-learn anders konfiguriert. Die Unterschiede betreffen unter anderem die Vorverarbeitung. Insbesondere wurden n-Gramm-Merkmale (1 bis 3) verwendet, statt Ein-Grammen („quadflor“). Setting-B ist sonst vergleichbar mit der Basiskonfiguration (logistische Regression, keine Konzeptmerkmale) von „quadflor“.

4.3 Ergebnisse

4.3.1 Quadflor

Von den vier „quadflor“-Konfigurationen lieferte die Basiskonfiguration mit Keywords[11] den höchsten f1-Wert (f1=0,394). Sie erzeugte im Mittel 3,16 Deskriptoren pro Dokument und konnte damit die anderen drei Parametrisierungen beim *recall* übertreffen, wo sie einen Wert von 0,335 erreichte. Die komplexeren Konfigurationen hatten im Schnitt nur ungefähr einen Deskriptor pro Werk zugewiesen, was sich in einem niedrigen *recall* (r<0,15) widerspiegelte. Mit Autoren-Keywords erreichte die komplexe Konfiguration einen f1-Wert von 0,228.

Die Keywords brachten bei den Basiskonfigurationen und bei den komplexeren Konfigurationen jeweils Vorteile, sowohl bei der *precision* (jeweils über sechs Prozentpunkte) als auch beim *recall* (über sechs Prozentpunkte bei der Basiskonfiguration, ca. ein Prozentpunkt bei der komplexeren Konfiguration).

Für jede der vier „quadflor“-Konfigurationen war die *precision* deutlich höher als der *recall*.

4.3.2 Setting-B

Auch diese Konfigurationen erzielten mit Autoren-Keywords deutlich bessere Ergebnisse als ohne sie. Die Konfiguration mit Autoren-Keywords ohne tf-idf Gewichtung erzielte insgesamt den höchsten f1-Score (f1=0,401; *precision*=0,584, *recall*=0,360). Knapp dahinter (f1=0,400) lag die Parametrisierung mit Autoren-Keywords und mit tf-idf. Die Hinzunahme des tf-idf Parameters führte zu einer höheren *precision*, aber einem niedrigeren *recall*. Die f1-Werte dieser beiden Einstellungen konnten gegenüber der anderen Konfiguration noch leicht verbessert werden.

Keywords führten auch bei den Konfigurationen von Setting-B zu Steigerungen, wiederum sowohl bei der *precision* (jeweils über acht Prozentpunkte) als auch beim *recall* (jeweils mehr als sechs Prozentpunkte). Auch bei diesen Versuchen lag die *precision* über dem *recall*.

5 Diskussion

Der Ansatz, von Autoren bereitgestellte Fremddaten zur Kontextanreicherung eines titelbasierten Modells zu benutzen, hat sich als vielversprechend erwiesen. Schon der einfache Ansatz hat die insgesamt besten Ergebnisse erzielen können. Sowohl *precision* als auch *recall* wurden durch Hinzunahme der Autoren-Keywords stets deutlich verbessert. Zum einen können demnach mehr Begriffe durch die zusätzlichen Terme der Autoren-Keywords zugewiesen werden. Zum anderen erfolgen die automatischen Verschlagwortungen genauer. Als Ursache vermuten wir den zusätzlich bereitgestellten Kontext.

[11] Weil Titel bei allen Konfigurationen verwendet wurden, beziehen sich Formulierungen wie „mit Keywords“ folgend auf die Eingabe „Titel+Keywords“.

Ohne die tf-idf Gewichtung wurden höhere f1-Werte erzielt als mit tf-idf. Wegen der Kürze der Titel und der Autoren-Keywords kommen die Terme nur relativ selten pro Dokument vor, weshalb die Gewichtung massgeblich von der inversen Dokumentenhäufigkeit bestimmt wird. Wir vermuten, dass sich die tf-idf Gewichtung besser für Volltexte eignet als für Titel.

Absolute Werte von *precision*, *recall* und f1-Mass anderer Publikationen zu automatischer Indexierung lassen sich im Allgemeinen nicht direkt vergleichen. Zu gross sind die Einflüsse von Zusammenstellung und Grösse der Trainings- und Testdatensätze sowie des Thesaurus. Im Manual von „quadflor“ wird beispielsweise berichtet, dass die Ergebnisse bei einer zufälligen Stichprobe von 65.000 Dokumenten im Vergleich zu einer Stichprobe mit ca. 62.000 Dokumenten, für die Volltextnutzungsrechte vorliegen, zurückfielen. Analog muss auch der Vergleich zur Konsistenz zwischen professionellen Indexierern gesehen werden.

6 Zusammenfassung und Ausblick

In diesem Beitrag haben wir einen Einblick in die Arbeit der Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW) im Bereich automatischer Indexierung gegeben. Wir haben einen direkten Ansatz benutzt, um von Autoren vergebene Schlagwörter in ein *multi-label classification framework* zu integrieren, das auf Titeldaten arbeitet. Unsere Experimente zeigen, dass die Methode bereits zu deutlichen Verbesserungen von *precision* und *recall* im Vergleich zu rein titelbasierten Ansätzen führen kann. Auf der Grundlage von Titel und Autoren-Keywords normierte inhaltsbeschreibende Daten zu erzeugen, erscheint somit als ein vielversprechender Ansatz, die Erschliessung grosser Datenmengen zu stemmen. Eine ausführlichere Analyse der Ergebnisse, insbesondere der Fehler des Systems, steht aus. Vor dem Hintergrund, dass Text- und Data-Mining Rechte für digitale Kollektionen oft Gegenstand schwieriger Verhandlungen mit Verlagen sind, liessen sich derart urheberrechtlich nicht geschützte Daten verwenden, um den Zugang zum Bestand allgemein und das Information Retrieval zu verbessern. Aktuell werden zudem im Rahmen des Mehr-Ebenen-Ansatzes des Erschliessungskonzepts (siehe Kap. Hintergrund) alternative Herangehensweisen für das automatische Indexierungssystem der ZBW sowie Verfahren zur Thesaurusanreicherung untersucht.

7 Disclaimer und Danksagung

Dieser Beitrag ist ein praxisbezogener Werkstattbericht aus dem Bibliothekskontext und kein technischer Beitrag mit neuartigen informatischen Methoden. Für ausführlichere Informationen zu den beschriebenen technischen Begriffen verweisen wir auf entsprechende Lehrbücher und Publikationen (Schlagwörter: Text Mining, Machine Learning, Computerlinguistik). Wir bedanken uns bei Alan Schelten, Lukas Galke, Dennis Brunsch und Florian Mai für die Bereitstellung der Software „quadflor“ und bei Tobias Rebholz für hilfreiche Kommentare und Anregungen.

Literatur

Alexander von Humboldt-Stiftung, Hrsg. (2009). *Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen. Beiträge zur Beurteilung von Forschungsleistungen*. URL: https://www.humboldt-foundation.de/pls/web/wt_show.text_page?p_text_id=1073898.

Auckland, M. (2012). *Re-skilling for Research. An investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers*. URL: http://www.rluk.ac.uk/wp-content/uploads/2014/02/RLUK-Re-skilling.pdf.

Bertram, J. (2005). *Einführung in die inhaltliche Erschließung*. Ergon Verlag.

Bornmann, L. und Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. In: *Journal of the Association for Information Science and Technology* 66.11, S. 2215–2222. DOI: 10.1002/asi.23329.

Faden, M. und Groß, T. (2011). *Automatische Sacherschließung an der ZBW – Status quo & Ausblick*. URL: http://files.dnb.de/pdf/petrus/automatische_sacherschliessung_zbw.pdf.

Ferber, R. (2005). Automated indexing with thesaurus descriptors: A co-occurrence based approach to multilingual retrieval. In: *Lecture Notes in Computer Science* 1324, S. 233–252.

Groß, T. und Faden, M. (2010). Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. In: *Bibliotheksdienst* 44.12, S. 1120–1135. URL: http://hdl.handle.net/11108/9.

Große-Bölting, G. (2015). *Vergleich verschiedener Verfahren zur automatischen Annotation von Dokumenten*. URL: https://www.kd.informatik.uni-kiel.de/en/bsc-msc-theses/accomplished-topics/msc-thesis-grosse-bolting.

Haller, J., Ripplinger, B. und Maas, D. (2000). *Automatische Indexierung von wirtschaftswissenschaftlichen Texten - ein Experiment*. URL: https://www.yumpu.com/de/document/view/3008682/automatische-indexierung-von-wirtschaftswissenschaftlichen-iai.

Haustein, S. und Peters, I. (2012). Using social bookmarks and tags as alternative indicators of journal content description. In: *First Monday* 17.11. DOI: 10.5210/fm.v17i11.4110.

Jurafsky, D. und Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2. Aufl. Pearson Education International. URL: http://www.cs.colorado.edu/~martin/slp.html.

Kempf, A. O. und Rebholz, T. (2016). 'Mixed Methods' Indexing. Building-Up a Multi-Level Infrastructure for Subject Indexing. In: *Subject Access: Unlimited Opportunities*. Classification & Indexing Satellite Meeting, IFLA, Columbus, Ohio. URL: http://hdl.handle.net/11108/259.

Klimpel, P. (2015). Eigentum an Metadaten? Urheberrechtliche Aspekte von Bestandsinformationen und ihre Freigabe. In: *Handbuch Kulturportale. Online-Angebote aus Kultur und Wissenschaft*. De Gruyter, S. 57–64.

Leininger, W. (2009). Wirtschaftswissenschaften. In: *Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen. Beiträge zur Beurteilung von Forschungsleistungen*. Hrsg. von Alexander von Humboldt-Stiftung, S. 67–68. URL: https://www.humboldt-foundation.de/pls/web/wt_show.text_page?p_text_id=1073898.

Medelyan, O. und Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In: *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2006)*, S. 296–297. DOI: 10.1145/1141753.1141819.

Rebholz, T. und Bahls, D. (2015). *Evidenzbasierte Begriffs- und Synonymerweiterung des STW*. 104. Deutscher Bibliothekartag in Nürnberg. URL: https://opus4.kobv.de/opus4-bib-info/frontdoor/index/index/docId/2498.

Schelten, A., Galke, L., Brunsch, D. und Mai, F. (2016). *lucidML. Technical Documentation*. Techn. Ber. URL: https://github.com/quadflor/Quadflor/blob/master/Documentation/technical_documentation.pdf.

Schröter, M. (2012). Fachreferat 2011 – Innenansichten eines komplexen Arbeitsfeldes. In: *Bibliothek, Forschung und Praxis* 36.1, S. 32–50. DOI: 10.1515/bfp-2012-0005.

Stumpf, G. (2015). *„Kerngeschäft" Sacherschließung in neuer Sicht. Was gezielte intellektuelle Arbeit und maschinelle Verfahren gemeinsam bewirken können*. URL: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:384-opus4-30027.

Wortmann, K., Groß, T. und Bahls, D. (2014). *Sacherschließung in der ZBW. Anwendung automatischer Verfahren – Werkstattbericht*. 103. Deutscher Bibliothekartag in Bremen. URL: https://opus4.kobv.de/opus4-bib-info/files/1571/2014_06_03_Beitrag2_automSacherschliessung_ZBW.pdf.