Eine Infrastruktur für „Big Metadata“

Magnus Pfeffer

Erfolgreich abgeschlossene Experimente des Autors mit Clustering-Verfahren zum Austausch von Erschließungsinformationen zeigen, welches Potenzial in der Anwendung automatischer Analyseverfahren auf bibliografische Daten steckt. Die zunehmende Bereitstellung nationaler und internationaler Katalogdaten als Open Data hat die Basis für solche Analysen nochmals verbreitert. Mit der Softwareplattform culturegraph.org existiert nun auch eine Basis, auf der Interessierte die vorgestellten Experimente nachvollziehen, modifizieren und an eigenen Daten ausprobieren können.

In experiments the author applied clustering methods to bibliographic data in order to exchange indexing information. The results were very encouraging. With a growing number of national and international libraries offering their data under open licenses, this and similar approaches could become increasingly powerful. With culturegraph.org there exists a software platform that allows interested parties to implement the experiments as well as to modify and adapt them to other data sets.

Einleitung

In den vergangenen Jahren hat es einen bedeutsamen Wandel in der Verfügbarkeit von bibliografischen Daten gegeben. Immer mehr Verantwortliche in den Bibliotheken haben erkannt, dass die Daten in ihren Katalogen und bibliografische Datensammlungen der Allgemeinheit zur Verfügung stehen sollten und diese im Sinne der Idee von *Open Data* unter permissiven Lizenzen in Standardformaten veröffentlicht. Allein in Deutschland sind zwischenzeitlich Daten aus allen Bibliotheksverbünden und der Deutschen Nationalbibliothek frei verfügbar oder ihre Freigabe ist zumindest angekündigt.[1] International ist das Bild ähnlich: Viele Nationalbibliografien sowie wichtige Normdateien sind zwischenzeitlich als Open Data verfügbar.[2]

In diesem Artikel soll an Beispielen das Potenzial aufgezeigt werden, das in Open Data aus Bibliotheken schlummert. Es ist dank der großen, frei verfügbaren Datenmenge nun ohne weiteres möglich, Methoden und Techniken der Analyse und automatischen Verarbeitung, die ursprünglich für andere Arten von Daten entwickelt wurden, auf bibliografische Daten anzuwenden. In vielen Fachgebieten und auch in der freien Wirtschaft spricht man schon länger von den Herausforderungen des „Big Data“: Riesige Datenmengen, die in Experimenten und Versuchen als Messdaten anfallen oder als Logdateien aus der Nutzung von Softwaresystemen entstehen, müssen nicht nur gespeichert, sondern auch verarbeitet, aufbereitet und letztlich ausgewertet werden. Auch Bibliotheken können von den Erfahrungen profitieren, die in diesen Bereichen gemacht werden.

Zur Illustration werden im nächsten Abschnitt zunächst einige Ergebnisse aus Projekten mit Metadaten vorgestellt. Im zweiten Teil werden einige der technischen Herausforderungen, die sich bei der Arbeit mit sehr großen Datenmengen ergeben, erläutert. Abschließend wird die Softwareplattform culturegraph.org vorgestellt, die unter der Federführung der Deutschen Nationalbibliothek entwickelt wird. Mit dieser Plattform können Interessierte mit vergleichsweise überschaubarem Aufwand selbst Analysen für große Datenmengen entwerfen und umsetzen.

Projekte mit Metadaten

Konsolidierung heterogener Erschließung über Kataloggrenzen

Durch die kooperative Erschließungspraxis kann es vorkommen, dass unterschiedliche Ausgaben eines Werks nicht einheitlich oder teilweise gar nicht inhaltlich erschlossen werden. Bei Werken, die in einer großen Zahl von Ausgaben erschienen sind, kann diese Heterogenität sehr ausgeprägt sein. Zur Illustration zwei Beispiele aus dem Katalog des Bibliotheksverbunds Bayern:

- „Herzfeld, Hans: Der erste Weltkrieg”
  18 Titelsätze, von denen nur 11 mit RSWK und 8 mit RVK erschlossen sind.

1 HeBIS: http://www.hebis.de/de/1ueber_uns/projekte/lod/lod_index.php
SWB: https://wiki.bsz-bw.de/doku.php?id=v-team:daten:openaccess:start
BVB: http://lod.b3kat.de/doc
HBZ: http://opendata.hbz-nrw.de/
DNB: http://www.dnb.de/lds
GBV: http://uri.gbv.de/downloads/

2 Als repräsentative Beispiele seien hier nur die Projekte der British Library (Nationalbibliografie, http://www.bl.uk/bibliographic/datafree.html) und der Library of Congress (Normdateien, http://id.loc.gov) genannt.

Katalog	Monografien gesamt	Monografien mit RVK	Monografien mit RSWK	Neu mit RVK erschlossen	Neu mit RSWK erschlossen
SWB	12.777.191	3.235.958	3.979.796	959.419	636.462
HeBIS	8.844.188	1.933.081	2.237.659	992.046	1.179.133

Tabelle 1: Ergebnisse des ersten Projektschritts

- „Tanenbaum, Andrew S.: Computer Networks" 44 Titelsätze, davon 38 mit RSWK und 31 mit RVK erschlossen. Es finden sich vier unterschiedliche RVK-Notationen aus Informatik, Wirtschaftswissenschaften und Soziologie.

In einem Projekt wurde untersucht, in welchem Umfang diese lückenhafte und inkonsistente Erschließung auftritt und ob sich durch eine geeignete Gruppierung der Titel ein automatischer Abgleich unter den unterschiedlichen Ausgaben eines Werkes realisieren lässt. Im nächsten Schritt wurde dieser Abgleich über mehrere Kataloge durchgeführt, so dass Erschließungsinformationen auch unter diesen ausgetauscht werden konnten.

Zur Konsolidierung müssen unterschiedliche Ausgaben eines Werkes sicher als solche erkannt und gruppiert werden. Dies betrifft z.B. unterschiedliche Auflagen, Ausgaben in verschiedenen Materialarten (z.B.

Print- und E-Book-Ausgabe) sowie Übersetzungen. Innerhalb der so entstehenden Cluster können dann die Erschließungsinformationen harmonisiert werden. Der gewählte Ansatz gruppiert Titel, bei denen Sachtitel und Untertitel exakt übereinstimmen und mindestens eine Person oder Körperschaft übereinstimmt. Dabei wird nicht nach Autoren, Herausgebern und sonstigen Beteiligten unterschieden. Durch die Einschränkung auf nur eine übereinstimmende Person/Körperschaft werden auch Ausgaben gruppiert, bei denen z.B. das Herausgeberteam erweitert wurde oder ein neuer Autor ein älteres Werk aktualisiert.

Durch die Vorgabe einer exakten Titelübereinstimmung werden Ausgaben desselben Werkes, bei denen sich zwischen den Auflagen der Sachtitel oder der Untertitel ändert, nicht erkannt.[3] Umgekehrt dürfte die Wahrscheinlichkeit, dass nicht zusammengehörende Titel gruppiert werden, aber sehr gering sein. In Fällen, bei denen Einheitssachtitel vergeben sind, wird auch dieser mit einbezogen. Damit können insbesondere Übersetzungen erkannt werden.

Die Ergebnisse sind sehr ermutigend. In einem ersten Testlauf mit Daten aus dem Südwestdeutschen Bibliotheksverbund (SWB) und dem Hessischen Bibliotheks- und Informationssystem (HeBIS) konnte eine sehr große Menge von Titeln identifiziert werden, die durch den Abgleich neu inhaltlich erschlossen werden konnten. Die genauen Zahlen sind in Tabelle 1 zusammengefasst.

Die ermittelten Erschließungsdaten wurden an die beiden beteiligten Verbundzentralen geliefert, welche die Schlagwörter und Notationen ganz oder in Auszügen zunächst in ihre Testdatenbanken einspielten. Weiterhin wurden im Rahmen eines Linked Data Projekts der UB Mannheim alle Titel und die gefundenen Cluster im Web präsentiert. Somit standen sowohl interessierten externen Experten als auch den Interessierten unter den Verbundteilnehmern die Daten für eine ausführliche Prüfung zur Verfügung. Die Arbeitsgruppen der Sacherschließungsexperten beider Verbünde haben die Ergebnisse dieser Prüfungen gesammelt und aufgrund der hohen Qualität der Daten ihre Einspielung in die Produktivsysteme empfohlen.[4]

In einem weiteren Projektschritt wurden Daten aus dem Hochschulbibliothekszentrum Nordrhein-West-

3 Es wäre zu prüfen, ob der Algorithmus durch den Einbezug von Informationen aus Fußnoten sowie eines Unschärfefaktors bei den Zusätzen zum Sachtitel noch verbessert werden kann.

4 Vgl. Protokoll der 20. Sitzung der AG Sacherschließung im HeBIS-Verbund vom 8.11.2010 (online unter http://www.hebis.de/de/1publikationen/protokolle/pdf/ag_sacherschliessung/10-11-08.pdf) und Protokoll der Sitzung der AG Sacherschließung des SWB vom 20.01.2011 (online unter http://verbund-swop.bsz-bw.de/volltexte/2011/316/pdf/AGSE110120.pdf)

falen (hbz) und dem Bibliotheksverbund Bayern (BVB) der Datenbasis hinzugefügt. Ziel war neben der Ermittlung der erschließbaren Titel in den beiden neu hinzugekommenen Verbundkatalogen auch die Beantwortung der Frage, inwieweit die in der ersten Projektphase berechneten Ergebnisse durch weitere Daten noch verbessert werden. Die wesentlichen Zahlen sind in Tabelle 2 zusammengefasst. Sie belegen den großen Erfolg: Wieder fanden sich mehrere Millionen Titel, die neu erschlossen werden konnten, und auch für die bereits bearbeiteten Kataloge ergab sich noch ein signifikanter Zugewinn.[5]

Katalog	Monografien gesamt	Monografien mit RVK	Monografien mit RSWK	Neu mit RVK erschlossen	Neu mit RSWK erschlossen
SWB	13.330.743	4.217.226	4.083.113	581.780	957.275
HeBIS	8.844.188	1.933.081	2.237.659	1.097.992	1.308.581
HBZ	13.271.840	1.018.298	3.322.100	2.272.558	1.080.162
BVB	22.685.738	5.750.295	6.055.164	2.969.381	2.765.967

Tabelle 2: Ergebnisse des zweiten Projektschrittes

Generieren von Konkordanzen

Die gefundenen Cluster lassen sich aber nicht nur für die Erhöhung der Sacherschließungsrate benutzen, sondern auch für andere Zwecke auswerten: So lassen sich aus dem gemeinsamen Auftreten von Klassen oder Schlagwörtern aus unterschiedlichen Erschließungssystemen Zusammenhänge zwischen diesen herleiten und zu einer automatisch generierten Konkordanz aggregieren. Für zwei Systeme A und B werden dazu die Cluster ausgewählt, die nach beiden Systemen erschlossen wurden. Für jede Klasse aus A wird ermittelt, wie häufig sie *insgesamt* auftritt und wie häufig dabei *zusammen* mit Klassen aus B. Findet sich für eine Klasse aus A nur eine gemeinsam auftretende aus B, so impliziert dies eine enge Übereinstimmung zwischen den beiden. Tritt eine Klasse aus A mit mehreren Klassen aus B auf, impliziert es eine grobe Übereinstimmung (die Inhalte der Klasse aus A teilen sich in B auf mehrere Klassen auf). Durch den Vergleich der relativen Häufigkeit lassen sich in diesem Fall auch graduelle Unterschiede zwischen den Beziehungen ermitteln. Die grundsätzliche Validität dieses Ansatzes konnte in einer Untersuchung des gemeinsamen Auftretens von SWD-Schlagwörtern und Notationen der Regensburger Verbundklassifikation (RVK) in den Daten des SWB bereits gezeigt werden.[6] Es ist zu erwarten, dass durch eine Verbreiterung der Datenbasis und eine Konsolidierung durch das Clustering noch bessere Ergebnisse erzielt werden können.

Technische Umsetzung

Den sehr positiven Ergebnissen aus dem Austausch von Erschließungsinformationen stehen deutliche Schwächen in der konkreten technischen Umsetzung gegenüber. Die für dieses Forschungsprojekt entwickelten Programme sind hochgradig auf die vorliegenden Daten zugeschnitten und erfordern eine spezielle Systemumgebung. Sie sind wenig dokumentiert und nur in sehr geringem Umfang anpassbar. Dazu kommt, dass der Bedarf an Hauptspeicher und Rechenleistung mit dem Umfang der zu bearbeitenden Daten wächst. Bereits die Verarbeitung von Datensammlungen in der Größenordnung von 50 Millionen Einträgen setzt relativ leistungsstarke Hardware voraus und beschäftigt diese über mehrere Stunden.

Für einen dauerhaften produktiven Einsatz – egal ob auf der Ebene der Verbundkataloge oder in der Anwendung auf kleinere Datenmengen – sind die vorhandenen Programme somit nicht geeignet. Benötigt wird vielmehr eine solide technische Plattform, die in einer verbreiteten und plattformunabhängigen Programmiersprache geschrieben und vollständig dokumentiert ist. Auf einer solchen Basis kann eine gemeinschaftliche Weiterentwicklung der Verfahren stattfinden, und Interessierte könnten davon ausgehend eigene Anpassungen vornehmen. Darüber hinaus wäre es wünschenswert, umfangreiche Berechnungen auf mehrere Rechner zu verteilen zu können.

5 Dabei ist zu beachten, dass der Datenabzug des SWB im zweiten Projektschritt aktualisiert wurde und bereits einen Großteil der im ersten Lauf ermittelten RVK-Notationen enthielt. Dadurch fällt der absolute Zugewinn kleiner aus.

6 Vgl. Probstmeyer, Judith (2009): Analyse von maschinell generierten Korrelationen zwischen der Regensburger Verbundklassifikation (RVK) und der Schlagwortnormdatei (SWD). HdM Stuttgart, Bachelorarbeit. Online unter http://opus.bsz-bw.de/hdms/volltexte/2009/667/

Culturegraph.org

Das Hochschulbibliothekszentrum Nordrhein-Westfalen und die Deutsche Nationalbibliothek standen im Rahmen eines gemeinsamen Projekts zum Aufbau eines Resolving- und Lookup-Dienstes für bibliothekarische Identifier vor einem ähnlichen Problem. Auch in diesem Projekt sollten bibliografische Daten aus den deutschen Verbundkatalogen und dem Katalog der Deutschen Nationalbibliothek zusammengeführt und nach unterschiedlichen Vorgaben analysiert und gruppiert werden. Die Ergebnisse dieses und laufender Folgeprojekte werden auf der Webseite http://www.culturegraph.org/ präsentiert. Erfreulicherweise steht die dafür entwickelte Software unter einer

Open-Source Lizenz zur Verfügung[7] und kann von jedem Interessierten an individuelle Bedürfnisse angepasst und erweitert werden. Die Software wurde von Anfang an nicht als experimenteller Prototyp, sondern unter der Maßgabe entwickelt, modular aufgebaut und einfach anpassbar zu sein. Weiterhin sollten auch sehr große Datenmengen gespeichert und effizient verarbeitet werden können. Die in der Programmiersprache Java geschriebene Plattform stützt sich dafür auf bereits existierende freie Software wie Apache Hadoop, HBase und Lucene und ist gut dokumentiert. Durch den Einsatz von Hadoop, einem Framework zur Entwicklung von skalierbarer Software, ist es möglich, die Verarbeitungsschritte auf mehrere Rechner zu verteilen. Dabei können sowohl eigene Rechner als auch die Angebote von Dienstleistern[8] genutzt werden. Um Anpassungen und Erweiterungen weiter zu erleichtern, wurde für die Softwareplattform eine eigene Beschreibungssprache entwickelt: „Metamorph“ stellt Funktionen bereit, die typischerweise bei der Verarbeitung von bibliografischen Daten benötigt werden und erlaubt es so auch Nicht-Programmierern, eigene Ideen umzusetzen.

Ausblick

In einem gemeinsamen Projekt mit der Deutschen Nationalbibliothek wird der Autor Anfang 2013 das vorgestellte Clustering-Verfahren auf diese Plattform übertragen und erste Anwendungen implementieren. Fest geplant sind zum einen der Austausch und die Konsolidierung von Erschließungsinformationen aller Art und zum anderen die Generierung einer automatischen Konkordanz zwischen der Dewey Decimal Classification (DDC) und der Regensburger Verbundklassifikation (RVK). Die Erfahrungen bei der Entwicklungs- und Portierungsarbeit werden in die Dokumentation einfließen und sollen als Handreichung den Einstieg in die Nutzung der Plattform erleichtern.

Für den im März 2013 stattfindenden Bibliothekskongress in Leipzig ist ein Workshop mit dem Titel „Anwendung von Clustering-Verfahren zur Verbesserung und Analyse von Katalogdaten“ geplant, auf dem diese Erfahrungen vorgestellt werden und der als Forum für den Austausch zu weiteren Entwicklungen und Ideen dienen soll. Aus der Community gibt es bereits erste Anregungen: So ist der Austausch formaler Erschließungselemente, wie z.B. die Verknüpfung mit individualisierten Normsätzen, innerhalb der Cluster denkbar. Auch die Aufbereitung der Cluster zu Datensätzen für die im Rahmen von FRBR und RDA diskutierten Abstraktionsebene der „Werke“ ist grundsätzlich möglich.

Es ist meine Überzeugung, dass die Entwicklung erst am Anfang steht. Die Zukunft wird zeigen, welche weiteren ‚Schätze‘ aus den vorhandenen Daten gehoben werden können.

Prof. Magnus Pfeffer
Studiengang Bibliotheks- und Informationsmanagement
Hochschule der Medien
Wolframstraße 32
70191 Stuttgart
pfeffer@hdm-stuttgart.de

7 Lizenz ist die Apache License 2.0. Der Quellcode wird auf der SourceForge Webseite angeboten: http://sourceforge.net/projects/culturegraph/

8 An vielen Hochschulen stellen die Rechenzentren Zugänge zu Hadoop-Servern zur Verfügung. Ein kommerzieller Anbieter ist Amazon mit dem Dienst „Elastic Map Reduce (EMR)“: http://aws.amazon.com/de/elasticmapreduce/