

Ulrike Junger

Möglichkeiten und Probleme automatischer Erschließungsverfahren in Bibliotheken

Bericht vom KASCADE-Workshop in der Universitäts- und Landesbibliothek Düsseldorf



Seit mehreren Jahren beschäftigt sich die Universitäts- und Landesbibliothek Düsseldorf in Zusammenarbeit mit der Fachrichtung Informationswissenschaft der Universität Saarbrücken mit den Möglichkeiten der automatischen Erschließung bibliothekarischer Daten. Die Deutsche Forschungsgemeinschaft hat zu diesem Themenkomplex mehrere aufeinanderfolgende Projekte an der ULB Düsseldorf gefördert.

Im Rahmen des ersten Projekts MILOS I (Maschinelle Indexierung zur erweiterten Literaturschließung in Online-Publikums-Katalogen) wurde ein Verfahren zur automatischen Indexierung bibliothekarischer Titeldaten entwickelt. Die Grundlage bildete das Saarbrücker Indexierungssystem IDX, ein wörterbuchbasiertes Verfahren, das u.a. die Ermittlung von Grundformen, Zerlegung von Komposita und die Erkennung von Mehrwortgruppen leistet. Anhand von 50 000 Titeldatensätzen der ULB Düsseldorf, die mit MILOS indexiert worden waren, wurde ein Retrievaltest durchgeführt, der gegenüber einer reinen Stichwortsuche sowohl eine Verbesserung des Recall wie auch der Präzision erbrachten¹.

Diese positiven Ergebnisse bildeten den Ausgangspunkt für das Nachfolgeprojekt MILOS II, das zum Ziel hatte, die Brauchbarkeit der automatischen Indexierung für die verbale Inhaltserschließung zu untersuchen und mit intellektueller Beschlagwortung zu vergleichen. Dafür wurde das Indexierungssystem durch die Integration der Sachschlagwörter aus der Schlagwortnormdatei (SWD) um eine semantische Komponente erweitert. Ein neuerlicher, diesmal von Angehörigen des Fachbereichs Bibliotheks- und Informationswesen der Fachhochschule Köln durchgeführter Retrievaltest, mit einem großen Pool von Titeldaten der Deutschen Bibliothek, erbrachte aus Sicht des Projektes wiederum erfreuliche Ergebnisse². Insbesondere die Nutzung der von der SWD bereitgestellten Begriffsrelationen (Synonyme, hierarchische Beziehungen) erbrachte eine Verbesserung der Suchmöglichkeiten. Als Konsequenz aus den Ergebnissen des Projekts stellte die ULB Düsseldorf die intellektuelle verbale Sacherschließung ein und setzt das MILOS-Indexierungssystem jetzt im Routinebetrieb ein.

Dennoch gaben die Erfahrungen aus MILOS I und II den Projektpartnern Anlaß, in einem weiteren Folgeprojekt einige bis dahin ungelöste Fragestellungen zu bearbeiten. Dazu gehören das Homonymenproblem bei automatisch generierten Indexdaten, der bisher fehlende Einsatz fortgeschrittener Retrievaltechniken und die unzureichende textliche Substanz, die bibliothekarische Titelaufnahmen als Grundlage für eine automatische Indexierung aufweisen.

Im Rahmen dieses dritten Projektes mit dem Namen KASCADE (Katalogerweiterung durch Scanning und

automatische Dokumenterschließung) veranstaltete die ULB Düsseldorf am 19. November 1998 einen eintägigen Workshop, um in vier Vorträgen die bisher erarbeiteten Ergebnisse der bibliothekarischen Öffentlichkeit vorzustellen und mit den Teilnehmern Probleme und Einsatzmöglichkeiten solcher Verfahren zu diskutieren³. Der Düsseldorfer Projektleiter Klaus Lepsky stellte im ersten seiner beiden Vorträge den bisherigen Verlauf des KASCADE-Projekts vor. Folgende Ziele bildeten dabei die Arbeitsgrundlage:

- Anreicherung von Titeldaten, um eine verbesserte Erschließung zu ermöglichen;
- Entwicklung eines Indexierungsverfahrens, das eine Selektion und Gewichtung von Deskriptoren erlaubt (SELIX);
- die Entwicklung eines Systems, das sog. Themen-Aspekt-Beziehungen identifizieren kann (THEAS);
- Durchführung eines neuerlichen Retrievaltests.

Ausgehend von der Feststellung, daß bibliothekarische Titelaufnahmen für eine tiefergehende maschinelle Erschließung nicht ausreichend sind, wurde eine Anreicherung dieser Daten um weitere dokumentbezogene Elemente durchgeführt. Konkret handelte es sich dabei um einen Bestand von 3000 rechtswissenschaftliche Titeln, deren Inhaltsverzeichnisse gescannt, die resultierenden Daten mit einer OCR-Software bearbeitet und in einer Datenbank abgelegt wurden. Diese Vorgehensweise erwies sich entgegen den Erwartungen als sehr problematisch und so zeitaufwendig, daß ein Einsatz als Routineverfahren nicht in Frage kommt. Bedauerlicherweise gibt es jedoch derzeit keine Alternativen zur Generierung derartiger Daten.

In einem weiteren Projektabschnitt wurden die angereicherten Titeldaten dann zweimal indexiert: die erste Indexierung beinhaltete die Bearbeitung der Datenbasis mit der MILOS-Software. In einer sog. Gewichtungsindeizierung wurden die ermittelten Indexate dann mit dem Programm SELIX bearbeitet, so daß am Ende für jedes Dokument eine Liste mit gewichteten Deskriptoren bereitstand. Um im Retrievaltest das SELIX-Verfah-

- 1 Lepsky, K., J. Siepmann u. A. Zimmermann: Automatische Indexierung für Online-Kataloge: Ergebnisse eines Retrievaltests. In: Zeitschrift für Bibliothekswesen und Bibliographie, 43 (1996), H. 1, S. 47-56.
- 2 Gödert, W., M. Liebig: Maschinelle Indexierung auf dem Prüfstand: Ergebnisse eines Retrievaltests zum MILOS II Projekt. In: Bibliotheksdienst 31 (1997), H. 1, S. 59-68.
- 3 Weitere Informationen findet man auf der WWW-Seite des KASCADE-Projekts unter http://www.uni-duesseldorf.de/WWW/ulb/kas_home.htm, sowie bei Lepsky, K. u. H.H. Zimmermann: Katalogerweiterung durch Scanning und Automatische Dokumenterschließung. Das DFG-Projekt KASCADE. In: ABI-Technik 18 (1998), H. 1, S. 56-60.

ren mit dem bisherigen MILOS-Verfahren vergleichen zu können, wurde eine zweite Indexierung der Rohdaten mit MILOS durchgeführt.

Über die Durchführung und die Ergebnisse dieses Retrievaltests berichtete Klaus Lepsky in seinem zweiten Vortrag. Die unabhängig von MILOS entwickelten Teilverfahren SELIX und THEAS wurden von den Saarbrücker Projektpartnern vorgestellt (siehe unten).

Ein aufwendiger Retrievaltest war bereits Bestandteil der beiden MILOS-Projekte. Während bei MILOS II der Test von Personen durchgeführt wurde, die nicht am Projekt beteiligt waren, unterzog sich das Düsseldorfer Projektteam diesmal selbst der Mühe. Den Ausgangspunkt bildete die bereits erwähnte Datenbank mit 3000 Dokumenten aus dem Fachgebiet Jura. Um möglichst realistische Bedingungen zu erreichen, wurden Mitarbeiter der juristischen Fakultät der Universität Düsseldorf um die Formulierung von Suchfragen gebeten. Mit 60 solcher Anfragen, die eine breite thematische Streuung, unterschiedliche Spezifität und Komplexität in der Formulierung aufwiesen, wurden nun eine Reihe von Suchläufen durchgeführt, wobei jeweils auf eine unterschiedliche Kombination von Indices zur Bildung der Ergebnismengen zugegriffen wurde (Sachtitel, Schlagwörter, Inhaltsverzeichnisse, ungewichtete versus gewichtete Deskriptoren, verschiedene Grenzwerte für die Einbeziehung gewichteter Deskriptoren). Die jeweils erzielten Treffermengen wurden wiederum den Mitarbeitern des Fachbereichs Jura zur Relevanzbeurteilung vorgelegt. Die Ergebnisse hinsichtlich Recall und Präzision bestätigten einerseits die durch vorangegangene Tests geformten Erwartungen der Projektgruppe, daß die Indexierung mit MILOS einen deutlich verbesserten Recall bei nur wenig verminderter Präzision erbringt. Ebenfalls bestätigt wurde die Hypothese, daß eine Anreicherung der Titeldaten um Indexate aus Inhaltsverzeichnissen den Recall verbessert. Hier leidet jedoch die Präzision. Andererseits brachte der Test einige Überraschungen. Das SELIX-Verfahren zur Gewichtung von Indexaten erbrachte weder für die Quantität noch für die Qualität der Treffermengen die erwarteten Effekte. Der Recall lag bei allen getesteten Cut-Off-Werten unter dem Ergebnis der MILOS-Indexierung, ebensowenig konnte die durchschnittliche Präzision nennenswert verbessert werden. Als Ergebnis bleibt festzuhalten, daß das SELIX-Verfahren in der bisherigen Form nicht für eine breitere Anwendung geeignet ist.

Zu Beginn des zweiten Teils des Workshops stellte Hubert Hüther aus Saarbrücken die Bestandteile des SELIX-Verfahrens vor. Ausgehend von einer Reihe dokument- wie bestandsbezogener Parameter, beispielsweise Häufigkeit eines Begriffs in einem Dokument oder der gesamten Datenbasis, werden Gewichte für die einzelnen Indexate berechnet, die als Grundlage für ein Cut-Off-Verfahren dienen. Der zugrunde liegenden Ansatz geht davon aus, daß ein Begriff dann für ein Dokument relevant ist, wenn er in diesem Dokument überdurchschnittlich häufig vorkommt. Eine Programmroutine aus Indexierungs-, Sortier- und Gewichtsläufen führt zu einer Menge an gewichteten Deskriptoren, auf die beim Information Retrieval zurückgegriffen wird.

In der sich anschließenden Diskussion wurde der kombinierte Einsatz von MILOS und SELIX kritisch in Frage gestellt. Durch die MILOS-Indexierung werden Komposita und Mehrwortgruppen in ihre Bestandteile zerlegt.

Dies könnte in SELIX jedoch zur Folge haben, daß die Einzelbegriffe aufgrund einer erhöhten relativen Häufigkeit nur ein geringes Gewicht erhalten. Ein möglicher Ausweg wird darin gesehen, solche komplexen Begriffe unabhängig von Gewichten in den Index aufzunehmen und sogar bevorzugt für das Information Retrieval zu nutzen.

Im letzten Vortrag des Workshops stellte Harald H. Zimmermann, ebenfalls von der Universität Saarbrücken, seine Konzeption einer Themen-Aspekt-Analyse (THEAS) vor. Dahinter steckt nicht, wie man auf den ersten Blick vermuten könnte, die Verknüpfung von verbaler und klassifikatorisch-systematischer Information, wie sie in Form von Notationen in bibliothekarischen Titelaufnahmen enthalten ist. Es wird vielmehr versucht, ein sog. semantisches Bezugssystem aufzubauen, das die Identifikation von Dokumenttypen, die Zuordnung von Dokumenten zu möglichen Zielgruppen sowie die eigentliche Themen-Aspekt-Identifizierung erlaubt. Dies soll anhand einer Sammlung und Analyse nominaler Strukturen (Komposita und Mehrwortgruppen aus Titeln und Untertiteln) erfolgen. Bei der Wendung „Auswirkungen des Waldsterbens“ würde „Waldsterben“ als Thema, „Auswirkungen“ dagegen als Aspekt identifiziert.

Im Rahmen des Teilprojekts THEAS wurde bisher jedoch lediglich ein Modell erarbeitet. Praktische Anwendungen oder ein Test in der KASCADE-Datenbank lagen zum Zeitpunkt des Workshops nicht vor.

In der lebhaften Diskussion, die sich an die Vorträge angeschlossen, wurden zunächst die Grenzen des KASCADE-Projektes angesprochen, die auch die Aussagekraft des Retrievaltests relativieren. Die Beschränkung auf eine vergleichsweise kleine Anzahl untersuchter Datensätze sowie vor allem die Beschränkung auf ein Fachgebiet lassen die Übertragbarkeit auf große, thematisch heterogene Datenbestände fraglich erscheinen. Ein weiteres Problem dürfte für große Katalogdatenbanken auch dadurch entstehen, daß durch die laufende Indexierung die Performance unzumutbar eingeschränkt wird. Dies wird insbesondere dann der Fall sein, wenn zur Disambiguierung von Indexaten der Kontext herangezogen werden soll. Das Performance-Problem dürfte sich auch bei einem Einsatz der Softwarekomponente SELIX stellen, da mit dem Hinzukommen neuer Einträge in die Datenbank jedesmal aufs neue die Gewichte der Indexbegriffe berechnet werden müßten. Nach Aussagen der Projektmitarbeiter ist dies ein Prozeß, der bereits für die Testdatenmenge von 3000 Sätzen mehrere Stunden in Anspruch nahm. Für den laufenden Betrieb einer umfangreichen Katalogdatenbank ist ein solches Vorgehen undenkbar.

Eine wesentliche, von Klaus Lepsky explizit benannte Schwachstelle der MILOS-Projekte ist im Verlauf des KASCADE-Projektes nicht beseitigt worden. Während das Problem der unzureichenden Textbasis für die Indexierung durch die Einbeziehung der Inhaltsverzeichnisse mindestens im engen Projektrahmen gelöst werden konnte, und durch den Einsatz von SELIX ein Ranking im Ansatz versucht wurde, blieb das Problem der fehlenden Bedeutungsdifferenzierung bei der automatischen Indexierung unbearbeitet bzw. wurde durch die Beschränkung der Datenbasis auf das Fachgebiet Jura ausgeblendet. Der im Teilprojekt THEAS verfolgte Ansatz scheint überaus kompliziert zu sein und darüber

hinaus für eine echte thematische Zuordnung von Dokumenten wenig auszutragen.

Hier stellt sich die Frage, warum die Düsseldorfer Projektgruppe bisher systematisch-klassifikatorische Sacherschließungsdaten nicht in ihre Projektkonzeptionen mit einbezogen hat. Derartige Information liegt in bibliothekarischen Titelaufnahmen in großem Umfang vor, teilweise – auf einen Titel bezogen – sogar mehrfach, wird jedoch für das Information Retrieval bisher nicht ausreichend genutzt. Hier läge die vermutlich einfachste Möglichkeit, eine Disambiguierung verbaler Indexate zu erzielen.

Die Teilnehmerrunde beschäftigte sich schließlich auch mit der Frage nach Einsatzmöglichkeiten für MILOS. Wie bereits erwähnt, setzt die ULB Düsseldorf MILOS routinemäßig für die Erschließung des laufenden Bestandszuwachses ein, ein Verfahren, das bisher bei anderen Bibliotheken auf wenig Gegenliebe gestoßen ist. Bibliothekarische Informationssysteme und -datenbanken bieten jedoch eine zunehmende Menge an Daten an, für die eine intellektuelle Sacherschließung nicht existiert und auch niemals geleistet werden kann. Beispiele hierfür sind Zeitschriftendatenbanken oder Titeldaten der Altbestandserschließung. Eine Anreicherung solcher Daten mit zusätzlichen Indexaten könnte für das Retrieval einen Gewinn bringen, ohne damit die intellektuelle Sacherschließung gleich in Frage zu stellen.

Es ist sicherlich ein Verdienst der Düsseldorfer Projekte, das Thema automatische Erschließung bibliothekarischer Daten auf die Tagesordnung gebracht zu haben, und es ist zu bedauern, daß bisher kein Versuch zur Replikation oder Reanalyse der im Rahmen von MILOS gewonnenen Ergebnisse unternommen wurde.

Man kommt jedoch an einer Feststellung nicht so ohne weiteres vorbei: auch Verfahren zur automatischen Erschließung setzen umfangreiche intellektuelle Vorarbeiten voraus. Im Falle von MILOS/KASCADE betrifft dies

die SWD, die mit hohem Aufwand gepflegt wird. Es darf die Vermutung geäußert werden, daß die Qualität solcher intellektueller (Vor)Arbeit sich auch auf die Qualität der davon profitierenden automatischen Verfahren auswirkt. Dies läßt umgekehrt den Schluß zu, daß beispielsweise die Pflege von Normdateien nicht überflüssig ist, sondern vielleicht sogar noch verbessert werden sollte. Wünschenswert wären gezielte Untersuchungen zu der Frage, in welchem Verhältnis klassische bibliothekarische Sacherschließung und automatische Erschließungsverfahren künftig stehen sollen, welcher Datenbestand auf welche Weise am besten erschlossen wird, und dies alles unter Berücksichtigung der damit verbundenen Kosten und des daraus resultierenden Nutzens. Eine genaue Prüfung ist auch deshalb geboten, weil nicht immer der Augenschein ein verlässlicher Maßstab ist.

Ein wichtiges Instrument bilden in diesem Zusammenhang Retrievaltests, wie sie die Düsseldorfer Projektgruppe bereits mehrfach durchgeführt hat. Andere Bibliotheken und insbesondere Verbundsysteme, die über Datenbanken mit breiter thematischer Streuung verfügen, wären hier gefordert. Der mit solchen Tests verbundene Aufwand sollte kein Argument dagegen sein, ebensowenig die Tatsache, daß die Relevanzbeurteilung für Treffermengen ein methodisches Problem darstellt. Dem kann z.T. dadurch begegnet werden, daß alle Daten eines solchen Retrievaltests offengelegt werden, um die Nachvollziehbarkeit der Ergebnisse zu gewährleisten.

Anschrift der Autorin:

Ulrike Junger
Fachreferentin für Psychologie
Niedersächsische Staats- und Universitätsbibliothek
D-37070 Göttingen