KLAUS LEPSKY / HARALD H. ZIMMERMANN

Katalogerweiterung durch Scanning und automatische Dokumenterschließung

Ergebnisse des DFG-Projekts KASCADE

Stand: 18.05.2000

Erschienen in: Zeitschrift für Bibliothekswesen und Bibliographie 4/00, 305-316

Der Beitrag befasst sich mit den Zielen, Inhalten und Ergebnissen des von der DFG geförderten Projekts KASCADE (KAtalogerweiterung durch SCanning und Automatische Dokument-Erschließung). Für KASCADE wurden Katalogdaten aus dem Fachbereich Rechtswissenschaft um Inhaltsverzeichnisse angereichert. Die angereicherten Titeldaten wurden mit einem erweiterten MILOS-Verfahren automatisch indexiert sowie mit den beiden linguistisch und statistisch basierten Verfahren SELIX und THEAS zusätzlich erschlossen. In einem umfangreichen Retrievaltest wurden die Ergebnisse der automatischen Indexierung und Gewichtung untersucht.

1. Projektidee[1]

Der Gedanke, dass die konventionelle bibliothekarische Titelaufnahme, versehen mit einer ebenfalls konventionellen bibliothekarischen Inhaltserschließung, eine zweifelhafte Basis für den Sucherfolg in Online-Katalogen liefert, ist beileibe nicht neu. Spätestens seit der Untersuchung von Lancaster et a1.[2] ist klar, dass Vokabularunterstützung und Vokabularanreicherung dringend erforderlich sind, um aus einer Titelaufnahme ein hinreichend gut wieder zu findendes Dokument zu machen.

Beides, Vokabularunterstützung und Vokabularanreicherung, waren denn auch die vordringlichen Ziele der seit 1993 durchgeführten MILOS-Projekte.[3] In enger Zusammenarbeit zwischen der Universitäts- und Landesbibliothek Düsseldorf und der Fachrichtung Informationswissenschaft der Universität des Saarlandes konnte mit Förderung der DFG ein System zur automatischen Indexierung für den Einsatz in Bibliotheken als routinemäßig einsetzbares Produktpaket entwickelt werden. Die durch Retrievaltests belegten Ergebnisse der MILOS-Projekte weisen deutliche Verbesserungen des Sucherfolgs für automatisch indexierte Titel- und Erschließungsdaten auf. Dies wird durch eine grammatikalische Vereinheitlichung des Vokabulars sowie die Berücksichtigung des semantischen Wortumfelds durch die Bereitstellung von Synonymverweisungen ermöglicht.[4]

Trotz dieser Erfolge war offenkundig, dass die Qualität einer automatischen Indexierung von bibliothekarischen Titeldaten gewissen Einschränkungen unterliegen muss, weil die Textbasis von Titelaufnahmen allgemein zu „dünn" ist. Nach Abschluss der MILOS-Projekte lag daher der Gedanke nahe, die Verfahren des Systems zur Sprachdatenverarbeitung auf der Basis umfangreicherer Textmengen weiter auszubauen, d. h. durch Dokumentanreicherung die Basis für den Sucherfolg zu verbessern.

Im dem wiederum von der DFG geförderten Projekt KASCADE sollten daher bibliothekarische Titelaufnahmen um zusätzliche Textinformationen aus Inhaltsverzeichnissen, Sachregistern und ggf. Abstracts angereichert werden. Die so angereicherten Titeldaten sollten einer automatischen Indexierung unterzogen werden und, darauf aufbauend, mit einer im Projekt zu entwickelnden

statistischen Komponente gewichtet indexiert werden. Hintergrund dieses Vorhabens war die Überlegung, dass eine „nur" gleichordnende Indexierung mit allen Termen aus dem Anreicherungsprozess neue Erschließungsprobleme verursachen werde, weil zu viele nicht-relevante Begriffe zu „Deskriptoren" werden. Lösbar schien dieses Problem durch ein Gewichtungsverfahren, das geeignet ist, aus der Vielzahl der verfügbaren Begriffe die relevantesten zu identifizieren und diese „echten Deskriptoren" dann für die Suche bereitzustellen.

Darüber hinaus sollte im Rahmen der KASCADE-Erschließung eine sogenannte„Themen-Aspekt-Identifizierung" erfolgen, die neben einer Dokument-Typisierung eine unmittelbare Vokabularunterstützung des Retrievals durch die Generierung von kanonischen Wortformen erlaubt.

Wie bereits in den beiden MILOS-Projekten stand auch für KASCADE ein umfangreicher Retrievaltest am Ende des Projekts, mit dem die unterschiedlichen Erschließungsvarianten (intellektuell und automatisch) miteinander verglichen werden konnten. Die wesentlichen Arbeitsinhalte des KASCADE-Projekts waren daher:

- Anreicherung der Titeldaten um zusätzliche inhaltsrelevante Daten aus Inhaltsverzeichnissen, Sachregistern und ggf. Abstracts. Die Dokumente für die KASCADE-Testdatenbank stammten aus dem Bestand des Faches Jura der Universitäts- und Landesbibliothek Düsseldorf. Es wurden nur deutschsprachige Dokumente bearbeitet. Die Daten wurden gescannt und mittels Zeichenerkennung (OCR) in Textdateien umgewandelt.
- Entwicklung einer gewichtenden automatischen Indexierung (SELIX).
- Entwicklung einer Systemkomponente zur Themen-Aspekt-Identifizierung und zur Erzeugung kanonischer Wortformen (THEAS).
- Durchführung eines Retrievaltests.

2. Projektarbeiten und Projektergebnisse

2.1 Dokumentanreicherung, Scanning und OCR-Texterkennung

Die KASCADE-Testdatenbank wurde an der Universitäts- und Landesbibliothek Düsseldorf durch Scanning von inhaltsrelevanten Daten aus 3.000 wissenschaftlichen Monografien aus dem Bestand des Faches Jura aufgebaut. Scanning und OCR-Texterkennung erfolgten mit der für die Projektzwecke angepassten Software newsWorks der Firma CCS, Hamburg.

Trotz sehr gut arbeitender Hard- und Software nahmen die Scanningarbeiten durch die studentischen Hilfskräfte erhebliche Zeit des Projekts in Anspruch (insgesamt ca. 15 Monate), sodass die vollständige Testdatenbank erst zu einem relativ späten Zeitpunkt zur Verfügung stand.

Die Ergebnisse des Scanning-Prozesses waren für die Inhaltsverzeichnisse zufrieden stellend, Abstracts fanden sich in der ausgewählten Literatur zu selten, Sachregister erwiesen sich als zu heterogen und waren nur in einer Teilmenge der Titel enthalten. Die vollständigen Ergebnisdaten des Scan-Prozesses - Scan als Grafik und Scan als Text - wurden in einer relationalen Datenbank (MS-Access) abgelegt, um eine einfache Verknüpfung mit den Ergebnisdaten der automatischen Erschließung und mit den zugehörigen Katalogdaten des Düsseldorfer Katalogs zu ermöglichen.

Die Anreicherung der Dokumente über Scanning erforderte dessen Einbindung in den Geschäftsablauf der Bibliothek. Dies führte im Laufe des Projekts zu einer Reihe von praktischen Problemen und Verzögerungen, die das ursprünglich formulierte Mengenziel (12.000 Dokumente) schon recht bald als illusorisch erscheinen ließen. Als Fazit aus diesen Erfahrungen darf fest-

gehalten werden, dass der für KASCADE gewählte Weg der Dokumentanreicherung über Scanning für zukünftige Modelle nicht notwendig die Methode der Wahl ist. Die Festlegung der projektrelevanten Daten als Kollektion von Büchern aus dem Bestand der Bibliothek ließ attraktivere Alternativen wie Fremddatenübernahme oder Kauf der gescannten Daten allerdings auch nicht zu.

2.2 Erschließung der Datenbasis

2.2.1 Selektive automatische Indexierung - SELIX

Wesentlicher Schwerpunkt der Projektarbeiten war die Software-Entwicklung für die selektive automatische Indexierung (SELIX), mit der über eine gewichtende Indexierung eine qualifizierte Deskriptorauswahl aus der Vielzahl gescannter Informationen erfolgen sollte. Erforderlich war die Neuentwicklung eines Gewichtungsalgorithmus sowie eine Einbettung des neuen Verfahrens in den bereits etablierten Indexierungsprozess mit MILOS.

SELIX setzt auf der computerlinguistischen Grundindexierung mit MILOS auf - grammatikalische Vereinheitlichung durch Grundformermittlung und Zerlegung von Komposita - und berechnet auf statistischem Wege Termgewichte für alle Begriffe aus Sachtiteln, intellektueller Verschlagwortung und gescannten Inhaltsverzeichnissen.

Dazu wurden verschiedene aus der Literatur bekannte statistische Gewichtungsverfahren - u.a. von Salton und Robertson - herangezogen und auf die vorgegebene Thematik adaptiert.[5] Im Unterschied zu diesen und anderen „Standard-Gewichtungsfunktionen" sollten im Rahmen der SELIX-Gewichtung jedoch auch bestimmte selten vorkommende Grundformen (komplexe Mehrwortgruppen oder Komposita) „statistisch gesehen" eine Chance erhalten. Die für SELIX neu entwickelte Gewichtungsfunktion benutzt daher die folgenden Parameter: [6]

- Länge der Dokumentenkollektion bezogen auf die erfassten Wortformen, Anzahl der Dokumente in der Kollektion, Häufigkeit einer Grundform in einem Dokument, Dokumentlänge bezogen auf die erfassten Wortformen, Anzahl der erfassten verschiedenen Grundformen in einem Dokument, Anzahl der Dokumente, die eine bestimmte Grundform enthalten, Anzahl des Vorkommens einer bestimmten Grundform in der Dokumentenkollektion, Länge einer bestimmten Grundform.

Auf der Basis dieser Parameter werden von SELIX drei Teilgewichte berechnet:

- Kollektionsgewicht

  Das Kollektionsgewicht gibt an, ob eine Grundform für eine gesamte Dokumentenkollektion als Indexterm geeignet ist.

- Dokumentgewicht

  Das Dokument gibt an, ob eine Grundform für ein Dokument als Indexterm geeignet ist.

- Längengewicht

  Das Längengewicht berücksichtigt die Länge eines Terms und ist damit unabhängig von Dokument und Kollektion.

Das Gesamtgewicht eines Deskriptors ergibt sich aus der Addition der drei Teilgewichte, wobei die Gewichtung der drei Teilgewichte untereinander durch Faktoren verändert werden kann. Das

Ergebnis der SELIX-Gewichtung wird als Termgewicht zu jedem Deskriptor geschrieben und liefert die Grundlage für die spätere Selektion der Deskriptoren.

Ausgehend von dem Gedanken, dass die Vielzahl der durch Scanning und linguistische Indexierung gewonnenen Begriffe nicht alle in gleichem Maße für das Retrieval relevant sein können, wurde die SELIX-Gewichtung als selektives Argument für die Auswahl von Deskriptoren verwendet (sog. CutOffPrinzip). Dabei wurde im Hinblick auf den späteren Retrievaltest mit CutOffWerten von 30, 100 und 200 gearbeitet, d. h. es wurden unterschiedliche Versionen der Testdatenbank aufgebaut, die die Dokumente mit den jeweils 30 bzw. 100 bzw. 200 „besten" Deskriptoren enthielten.

Mit dem Abschluss des Projekts steht das Indexierungssystem SELIX in einer stabilen Version unter Win-9 zur Verfügung. Die dem MILOS-Eingangsformat angeglichenen SELIX-Quelldaten werden in einer Abfolge von fünf Programmmodulen (KaskoGew, SortW, KasHfk, KasGew, SortW, KasSto) quantitativ analysiert, sortiert und ausgegeben. Zwischenergebnisse der Indexierung liegen als Datenbank bzw. Wörterbuch vor, die Endergebnisse im standardisierten MILOS-Ausgabeformat.

Die über das CutOff-Verfahren ausgewiesenen Deskriptoren wurden zusammen mit den Sachtiteln und ggf. Schlagwörtern der Dokumente zusätzlich mit der semantischen Komponente von MILOS II automatisch indexiert und so um evtl. lexikalisierte Synonyme ergänzt. Abschließend wurden die Titelaufnahmen, die Scanningresultate und alle Indexierungsergebnisse über die relationale Datenbank miteinander verknüpft und für den Retrievaltest abgelegt.

2.2.2 Themen-Aspekt-Analyse – THEAS

Durch die Anreicherung von Dokumenten sowie die Verfügbarkeit von elektronischen Volltexten steigt die Zahl verbaler Sucheinstiege über Stichwörter im Retrieval enorm an. Neben den o.a. selektiv wirkenden Maßnahmen auf der Vokabularebene über eine gewichtete Indexierung der Dokumente erschien die Entwicklung einer zusätzlichen klassifikatorischen Erschließungskomponente unter Retrievalgesichtspunkten sinnvoll und wünschenswert. Mit der Funktion einer so genannten „Themen-Aspekt-Identifizierung" wird versucht, eine Zuordnung des Dokuments zu (einer) möglichen Thematik(en) und möglichen Aspekten dieser Thematik automatisiert durchzuführen. Die Identifizierung erfolgt dabei auf der Grundlage eines semantischen Bezugssystems, das anhand der maschinenlesbaren Datensammlung durch statistische Analyse und intellektuelle Markierungen spezifischer lexikalischer Einträge aufgebaut wird.

THEAS baut auf den Ergebnissen der Dokumenterschließung durch das System MILOS auf. Soweit verfügbar und „brauchbar", werden auch die Ergebnisse der selektiven Indexierung bzw. der selektiven Indexierung mit kontrolliertem Vokabular verwertet.

In einem ersten Schritt wurde versucht, anhand der Analyse der verfügbaren Materialien (aus dem Bereich Jura) ein Kategorien- und Merkmalsystem abzuleiten, das diesen Ansprüchen gerecht wird:

- Ein erster Ansatz war der Versuch, aus den Volltext-Daten eine Art „Zielgruppe" der Nutzung abzuleiten. Dazu wurden Wörter und Wortgruppen, die entsprechende Hinweise geben können, entsprechend markiert. Beispielsweise geben Wörter wie „ein Kommentar zu" oder „ein praktisches Handbuch", aber auch Wörter wie „...prinzip" Hinweise auf eine eher theoretische oder praxisorientierte bzw. auf Experten oder auch Laien ausgerichtete

Zielgruppe. So weisen „Fallstudien" ggf. auf Studierende der Rechtswissenschaft usf. Derartige Hinweise finden sich in der Regel im Titel, im Untertitel oder auch im Abstract.

- Ein weiteres - eher formales - Kriterium für eine Selektion ist der Dokumenttyp. Soweit dieser nicht aus den Formaldaten (z. B. Formal-Schlagwörter aus der SWD) zu erschließen ist - bzw. auch ergänzend dazu -, ergeben sich vielfach wiederum aus dem Titel oder Abstract entsprechende „verbale" Hinweise, etwa auf eine „Gesetzessammlung", eine „Festschrift", einen „Kommentar", ein „Wörterbuch" usf.
- Die „eigentliche" Aufgabe bestand jedoch in der Differenzierung von Themen- und Aspekt-Beziehungen. Hierzu wurde ein Inventar verbaler und struktureller Indikatoren entwickelt. So geben beispielsweise Elemente wie „neue Erkenntnis" oder „aktuelle Entwicklung" Hinweise auf die „Aktualität eines Themas" (immer relativ zum Publikationsjahr); Elemente wie „ein Vergleich zwischen" (x und y) stellen entsprechende „Themen-Beziehungen" her, „Lokalisierungen" wie „in Bayern" können ein Thema einschränken usf. Aber auch „lexikalisierte" Strukturen dienen der Differenzierung, etwa bei Komposita und Mehrwortbegriffen wie „Prinzipien des ..." bzw. „...prinzip".

Der hohe Entwicklungsaufwand für SELIX ließ die vollständige Realisierung des ursprünglichen THEAS-Konzeptes im Rahmen der Projektlaufzeit nicht zu. Aufgebaut werden konnte ein „erstes Inventar" von „Regeln", die für die Erschließung einer Zielgruppe und des möglichen Dokumenttyps herangezogen werden. Ferner wurde für den Titelteil der ausgewählten Datenmenge ein lexikalisches Grundinventar erstellt, das für gängige Strukturen die ThemenAspekt-Zuordnung ermöglicht. Daraus kann ein allgemeineres, „offenes" Regelsystem abgeleitet werden, das es nach entsprechender technischer Implementierung erlaubt, diese Zuordnungen weitgehend ohne lexikalische Vorkodierungen zu realisieren. Das Lexikon wird dann nur noch vorgeschaltet, um „Ausnahmen" zu identifizieren bzw. die Bausteine für eine Themen-Aspekt-Zuordnung zu ermitteln.

Prototypisch implementiert wurde jedoch eine Themen-Aspekt-Analyse für Komposita und nominale Mehrwortgruppen. Diese leistet über eine Identifizierung von Mehrwort-Stichwörtern und Komposita eine Abbildung von der „natürlichen Wortfolge" auf eine sogenannte „kanonische Wortfolge", die dann für das Retrieval bereitgestellt wird:

- Datenschutz => Schutz, Daten

- Schutz personenbezogener Daten => Daten, personenbezogene; Schutz von.

Für die prototypische Testdatenbank wurden die von THEAS generierten kanonischen Einträge in ein (elektronisches) Register eingetragen. Während des Retrievals ist dadurch der Zugriff auf Mehrwortgruppen und (Bestandteile von) Komposita über eine standardisierte Fassung möglich, die darüber hinaus den Kontext des ursprünglichen Eintrags wahrt. Erste Versuche mit der THEAS-Komponente verliefen viel versprechend.

2.3 Retrievaltest

Wie schon in den vorangegangenen MILOS-Projekten war auch für KASCADE die Evaluierung der Projektergebnisse im Rahmen eines Retrievaltests wesentliches Projektziel. Für den KASCADE-Test wurde das Freitext-Retrievalsystem askSam verwendet, in das die Testdatenbank mit 3.000 vollständig KASCADE-erschlossenen Dokumenten geladen wurde. Der Wahl

von askSam gingen Tests mit den Systemen BISMAS 2.0, CONTEXT-PC und MicroOPC voraus, wobei letztlich dem „klassischen" Freitextretrieval der Vorzug gegeben wurde.

Anders als in den vorausgegangenen MILOS-Retrievaltests erfolgten die Erstellung und Auswahl der insgesamt 73 Suchfragen wie auch die anschließende Bewertung der Retrievalergebnisse in Zusammenarbeit mit Mitarbeiterinnen und Mitarbeitern der Juristischen Fakultät der Heinrich-HeineUniversität Düsseldorf. Dadurch war gewährleistet, dass das Spektrum der ausgewählten Suchanfragen wie auch die Auswertung der Testergebnisse von Fachleuten durchgeführt wurden. Nur ein geringer Teil der Suchanfragen wurden vom Projektteam festgelegt, um eher technische Probleme des Retrievals ebenfalls berücksichtigen zu können. Folgende vier Index-Kategorien wurden im Test verglichen:

- MILOS-Indexrate von Sachtitel und Schlagwörtern,
- MILOS-Indexrate von Sachtitel, Schlagwörtern und Inhaltsverzeichnissen,
- SELIX-Indexrate mit den 30/100/200 höchstgewichteten Deskriptoren aus Sachtitel, Schlagwörtern und Inhaltsverzeichnissen,
- Freitext auf Sachtitel, Schlagwörter und Inhaltsverzeichnisse.

Lediglich 60 der 73 Suchfragen konnten für den Test verwendet werden, weil für 13 Suchfragen keine Treffer in der Testdatenbank nachgewiesen wurden. Da es für die Gestaltung der Suchfragen an die juristischen Fachleute keinerlei Vorgaben gab, waren die gewählten Themen breit gestreut. Dabei waren wiederum typisch: eine breite Streuung von engen und weiten Themen, eine hohe Zahl von (verknüpften) Komposita, ein Nebeneinander von Fachtermini und Allgemeinbegriffen und eine hohe Zahl von Fragen mit mehreren Begriffen.

Insgesamt wurden für 60 Suchanfragen 1.421 Dokumente gefunden, wovon wiederum 876 relevant waren. Die durchschnittliche Zahl der Wörter in den für die Suche bereitgestellten Kategorien schwankte erheblich: von 30 Wörtern in der SELIX-30-Kategorie (Cutoff) bis zu durchschnittlich 1.881 Wörtern in den Volltexten der Inhaltsverzeichnisse. Die durchschnittliche Länge der Suchbegriffe betrug 14,5 Zeichen. Einige wichtige Ergebnisse des Retrievaltests:

- Dokumentanreicherung verbessert den Retrievalerfolg deutlich. Der Recall von 0.06 für eine Suche mit Titel und Schlagwörtern stieg für eine Suche mit Titelstichwörtern, Schlagwörtern und Inhaltsverzeichnissen auf 0.54. Die Precision sank dabei gleichzeitig von 0.98 auf 0.75.
- Die MILOS-Indexierung verbessert den Recall bei akzeptabler Precision deutlich. Gegenüber der nicht indexierten Version der Suche mit Titelstichwörtern, Schlagwörtern und Inhaltsverzeichnissen stieg der Recall von 0.54 auf 0.92, wobei die Precision nur unerheblich auf 0.70 sank.
- Die Recall-Werte für alle SELIX-Suchvarianten (30/100/200) sind mit 0.24/0.57/0.77 zu niedrig. Dies bedeutet, dass durch die Verwendung des CutOffs auch relevante Deskriptoren ausgeschlossen werden.
- Die Precision-Werte für die SELIX-Indexierung sind im Durchschnitt nicht signifikant besser als die Werte für die MILOS- oder Freitextvariante.
- Bei 27 Fragen erzielte SELIX/KASCADE die höchste Precision, wobei 17 Fragen aus zwei und mehr Suchbegriffen bestanden. Für 10 KompositumFragen erzielte SELIX/KASCADE eine höhere oder gleiche Precision.

Insgesamt haben die Ergebnisse des Retrievaltests nicht ganz die Erwartungen der Projektgruppe bestätigt. Es hat sich - vereinfacht zusammengefasst - als nicht zweckmäßig erwiesen, die Gewichtungsfunktion als selektives Argument für die Auswahl von Deskriptoren zu verwenden (CutOff-Prinzip). Im Gegenteil: Die Retrievalergebnisse verbessern sich stetig mit der Zahl der zugelassenen Deskriptoren, um mit einer reinen MILOS-Indexierung der erweiterten Titeldaten ihren optimalen Wert zu erreichen.

Diese Beeinträchtigung des Recall - die in gewissem Umfang zu erwarten war - wird nicht durch eine deutliche Verbesserung der Precision-Werte aufgewogen. Andererseits erwiesen sich die Ergebnisse der KASCADE-Erschließung als immer dann der „einfachen" Erschließung überlegen, wenn die Suchanfragen besonders komplexer Natur waren.

Eine der Ursachen für diese Ergebnisse ist zweifellos die im Testpool vertretene hohe Zahl von Suchen mit verknüpften Sachverhalten. Bei der UND-Verknüpfung von zwei oder mehr Suchbegriffen führt die SELIX-Suche immer dann zu einem Null-Treffer-Ergebnis, wenn bereits einer der Suchbegriffe unterhalb des CutOff-Wertes liegt. Dies ist dann besonders wahrscheinlich, wenn eine verknüpfte Suche aus einem eher allgemeinen Begriff und einem (oder mehreren) speziellen Begriff(en) besteht, weil dann die allgemeinen Begriffe mit hoher Sicherheit „rausgewichtet" werden (Beispiel: Haustürwiderrufsgesetz UND Geschichte).

Nachdem die Testergebnisse vorlagen, wurde nach einer Lösung gesucht, die „beiden" Verfahren - MILOS wie KASCADE - gerecht werden könnte. Als Lösung bietet sich an, die Gewichtung nicht zur Vorselektion zu nutzen (also einen CutOff-Wert anzusetzen), sondern zur Sortierung der Ergebnisse nach Wahrscheinlichkeiten, d. h. für das Ranking der Ergebnismenge. Damit wird z. B. das (zuvor bestehende) Problem gelöst, dass bei einer Suche mit einer UND-Verknüpfung von Deskriptoren bei KASCADE kein Ergebnis geliefert werden konnte, wenn das Gewicht eines der Deskriptoren unterhalb der CutOff-Schwelle lag, während sich ein entsprechendes Element bei MILOS qualifizierte. Durch eine einfache „Ausmultiplikation" der Gewichte lässt sich eine Ordnung nach Wahrscheinlichkeiten der Ergebnisse erreichen. Zumindest in einem entsprechenden Nachtest hat sich gezeigt, dass sich die Gesamt-Ergebnisse deutlich verbessern lassen.

3. Zusammenfassung der Projektergebnisse

Die mit KASCADE angestrebte Entwicklung eines gewichtenden automatischen Indexierungsverfahrens für angereicherte bibliothekarische Titeldaten konnte in großen Teilen erfolgreich abgeschlossen werden:

- Das SELIX-Gewichtungsverfahren arbeitet stabil und liefert zuverlässige Ergebnisse auch für sprachliche Problemfälle (Berücksichtigung großer Wortlängen; Berücksichtigung von Teilwörtern von Komposita).
- Das modulare Programmsystem der MILOS- und KASCADE-Erschließungskomponenten konnte im Hinblick auf die Gesamtperformance innerhalb des Projekts deutlich verbessert werden.
- Das THEAS-Entwicklungskonzept konnte in dem wichtigen Teilbereich der Identifizierung von Themen-Aspekt-Beziehungen realisiert und die Ergebnisse in einem Prototypen getestet werden.
- Ein sehr umfangreicher Retrievaltest konnte unter fachlicher Beteiligung erfolgreich abgeschlossen werden. Die Ergebnisse des Retrievaltests erlauben einerseits wichtige Rück-

schlüsse für den Vergleich unterschiedlicher automatischer Erschließungsmethoden, andererseits führten die Erfahrungen aus dem Retrievaltest zu weiteren Verbesserungen des Gewichtungsverfahrens.

Insgesamt weisen die KASCADE-Ergebnisse eindeutig die Richtung, in der eine zukünftig zu entwickelnde ernst zu nehmende Komponente zur automatischen Inhaltserschließung zu suchen sein wird:

- Die Orientierung am textreichen Dokument bis hin zum Volltext ist unausweichlich, denn nur auf einer breiten Textgrundlage lassen sich Systeme entwickeln, die hinreichend zuverlässig automatisch erschließen. Für den bibliothekarischen Bereich bringt die Einbeziehung eines Inhaltsverzeichnisses bereits deutliche Qualitätsverbesserungen gegenüber dem Standardsuchweg auf reine Titel- und Erschließungsdaten.
- Je umfangreicher die zu erschließenden Dokumente, um so wesentlicher ist die Ergänzung der linguistischen Komponente um eine statistische Komponente, weil nur so eine Relevanz-Differenzierung auf der Ebene der Treffermenge möglich ist.
- Zur Realisierung einer zuverlässigen automatischen Erschließung ist eine qualitätssichernde automatische Indexierung wesentlich. Diese kann nur über den Einsatz von qualitativ hochwertigen elektronischen Wörterbüchern erzielt werden, für deren Erstellung und Verwendung die MILOSIndexierung um eine Kontext berücksichtigende Komponente erweitert werden sollte.
- Das Retrieval auf Volltexte und/oder umfangreich verbal erschlossene Dokumente erfordert terminologische Unterstützung, die auf der Indexierungsseite wie auf der Retrievalseite über den Einsatz eines semantisch basierten begrifflichen Netzes realisiert werden kann. Das THEAS-Konzept innerhalb von KASCADE leistet hier wichtige Vorarbeiten.

KASCADE und MILOS stellen wichtige Etappen auf dem Weg zu einer qualitativ hochwertigen textbasierten Dokumenterschließung dar. Die zukünftigen Entwicklungen müssen bzw. werden notwendig Fragen der Multilingualität (zumindest des multilingualen Zugangs zu den Textquellen) und der automatischen Klassierung mit einbeziehen.

Anmerkungen:

[1] Der vorliegende Bericht beschränkt sich auf eine kurze Zusammenfassung der wesentlichen Projektinhalte und -ergebnisse. Eine breit angelegte Darstellung des Projekts erscheint demnächst als Band 31 der Schriften der Universitäts- und Landesbibliothek Düsseldorf. Mehrere Beiträge des KASCADE-Workshops in der Universitäts- und Landesbibliothek vom November 1998 wie auch die KASCADE-Testdatenbanken sowie weitere Hintergrundinformationen stehen auf der KASCADE-Homepage zur Verfügung:

http://www.ub.uni-duesseldorf.de/projekte/kascade/kas_home.

Vgl. auch:

Lepsky, Klaus: DFG-Projekt KASCADE an der ULB Düsseldorf. In: ProLibris, 3, 1997, S. 136.

Klaus Lepsky, Harald H. Zimmermann: Katalogerweiterung durch Scanning und Automatische Dokumenterschließung: Das DFG-Projekt KASCADE. In: ABI-Technik, 18, 1998, H. 1, S. 56-60.

Lepsky, Klaus: Automatische Erschließung von Internet-Quellen: Möglichkeiten und Grenzen. In: Buch und Bibliothek 50, Heft 5, 1998, S. 336-340.

Lepsky, Klaus: Automatische Indexierung in der Inhaltserschließung. In: 7e Dag van her Document. 19 & 20 mei 1998. Congrescentrum De Reehorst, Ede. Groningen 1998. S. 12-20.

Lepsky, Klaus: KASCADEWorkshop in der Universitäts- und Landesbibliothek Düsseldorf. In: ProLibris, 4, 1999, H. 1, S. 4.

Lepsky, Klaus: Automatische Indexierung zur Erschließung deutschsprachiger Dokumente. In: nfd. 50, 1999, S. 325-330.

[2] Lancaster, F.W., T.H. Connell, N. Bishop u.a.: Identifying barriers to effective subject access in library catalogs. In: Library resources and technical services. 35 (1991), S. 377-391.

[3] Vgl. hierzu:

Lepsky, K.: Maschinelles Indexieren zur Verbesserung der sachlichen Suche im OPAC: DFG-Projekt an der Universitäts- und Landesbibliothek Düsseldorf. In: Bibliotheksdienst 28(1994), H. 8, S. 1234-1242.

Lepsky, K.: Automatisierung in der Sacherschließung: Maschinelles Indexieren von Titeldaten. In: 85. Deutscher Bibliothekartag in Göttingen 1995: Die Herausforderung der Bibliotheken durch elektronische Medien und neue Organisationsformen. Hrsg.: S. Wefers. Frankfurt: Klostermann 1996, S. 223-233. (Zeitschrift für Bibliothekswesen und Bibliographie: Sonderh. 63).

Lepsky, K.: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. In: Zukunft der Sacherschließung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995. Hrsg.: E. Niggemann u. K. Lepsky. Düsseldorf: Universitäts- und Landesbibliothek 1996, S. 13-36. (Schriften der Universitäts- und Landesbibliothek Düsseldorf; Bd. 25).

Lepsky, K.: Inhaltserschließung von bibliothekarischen Massendaten. In: Ressourcen nutzen für neue Aufgaben: 86. Deutscher Bibliothekartag in Erlangen 1996. Hrsg.: S. Wefers. Frankfurt a.M.: Klostermann 1997, S. 296-306. (Zeitschrift für Bibliothekswesen und Bibliographie: Sonderh. 66).

MILOS: Automatische Indexierung für Bibliotheken: Handbuch. Hrsg.: Softex GmbH Saarbrücken und Universitäts- und Landesbibliothek Düsseldorf. Stand: Juni 1996. Düsseldorf: Universitäts- und Landesbibliothek 1996. Homepage: http://www.uni-duesseldorf.de/WWW/ulb/mil_home.htm.

[4] Vgl. Winfried Gödert, Klaus Lepsky: Semantische Umfeldsuche im Information Retrieval. In: Zeitschrift für Bibliothekswesen und Bibliographie 45, Heft 4, 1998, S. 401-423.

[5] Vgl. Huang, Xiangji; Robertson, S. E.: Okapi Chinese text retrieval experiments at TREC-6. In: The Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Special Publication 500-240, S. 137-142. Online publication: http: //trec.nist.gov/ pubs/trec6/t6_proceedings.html.

Salton, G.: Automatic Text Processing, Reading, Mass. 1989.

[6] Vgl. Hubert Hüther: Selix im DFG-Projekt Kascade. In: Harald H. Zimmermann, Volker Schramm (eds.): Knowledge Management und Kommunikationssysteme, Proceedings des 6. Internationalen Symposiums für Informationswissenschaft (ISI #98) in Prag, UVK Universitätsverlag Konstanz, 1998, S. 397-403.