Entwicklung eines Verfahrens der automatischen Klassifizierung für Textdokumente aus dem Fachbereich Informatik mithilfe eines fachspezifischen Klassifikationssystems

Humboldt-Universität zu Berlin

Philosophische Fakultät I Institut für Bibliothekswissenschaft

Masterarbeit zum postgradualen Fernstudium der Bibliothekswissenschaften M.A. (LIS)

Anja Schaar

Matrikelnummer: 500983/10. Matrikel

Gutachter: Prof. Dr. Umstätter, Dr. Lindtner

Rostock, 16.05.2006

Inhaltsverzeichnis

Einleitung

Bibliotheken, Dokumentationseinrichtungen, sowie Firmen im Medien und Verlagswesen verfügen in zunehmendem Masse über eine große Menge an elektronisch erstellten Dokumenten, die erschlossen, systematisiert und wieder auffindbar gemacht werden sollen. Zum Zwecke der Strukturierung eines Dokumentenbestandes gibt es Verfahren, die aus dem Inhalt der vorhandenen Dokumente ein Begriffsnetz generieren, um so leichter im Bestand navigieren zu können. Diese Verfahren sind in einigen Unternehmen schon fester Bestandteil und haben dort erheblich zur Verbesserung der Strukturierung des Dokumentenbestandes und zum schnelleren Wiederauffinden von Online-Texten geführt [SCHE2005]. Sinn der Generierung von Begriffsnetzen ist die Strukturierung der vorhandenen Texte ohne die Vorgabe von fixen Begriffs- und Benennungssystemen. Die Ausgangsposition ist im Bibliotheksumfeld jedoch etwas anders. Im Bibliotheksumfeld existieren feste Klassifikationen, die angewandt werden müssen, unabhängig von zu klassifizierenden Texten. Die Klassifikationen sind meist nur sehr wenig mit weiteren Informationen angereichert. Weiterhin kommen in den meisten wissenschaftlichen Bibliotheken bibliothekarische Hilfsmittel zur Themeneingrenzung und Schlagwortvergabe zur Anwendung. Es existieren Dokumente, die Schlagworten und oder Klassifikationen zugeordnet werden sollen. Eine Zuordnung wird intellektuell vorgenommen, da der Stand der Technik in den Bibliotheken keine andere Vorgehensweise erlaubt.

In der vorliegenden Arbeit werden Möglichkeiten für eine Automatisierung untersucht, evaluiert und im Ansatz umgesetzt. Dabei wird bereits im Entwurf sichtbar, dass es keine rein automatische Klassifikation geben kann, da es immer einen Bruch zwischen der intellektuell erstellten Klassifikation und dem aus dem Texten generierten Informationen geben wird. Hier kann nur ein semiautomatisches Verfahren entstehen, das im besten Fall durch die Anwenderaktionen lernt und auf eine umfangreiche und fachspezifi-

vollen Verwertung der manuellen Tätigkeiten bei diesem Vorgang und deren sukzessiver Verringerung. Die Ergebnisse einer semiautomatischen Klassifikation, werden mit denen einer manuellen verglichen. Am Ende wird ein Ausblick auf Möglichkeiten und Grenzen der automatischen Klassifikation gegeben.

1 Szenario

Ausgangspunkt für die Betrachtungen ist die Veranschaulichung der Arbeit des Fachreferenten bei der Sacherschließung, da seine Arbeit des Klassifizierens automatisiert werden, bzw. zumindest eine Unterstützung dieses intellektuellen Prozesses erreicht werden soll. Die Darstellung geschieht aus der Sicht des Information Retrieval und Text Mining.

Normalerweise wird ein Fachreferent kein intensiver Leser der ihm vorliegenden Texte sein. Der Fachreferent erfasst formale und inhaltliche Elemente des Textes, wie z.B. Angaben zu Verfasser, Publikationsform, Titel, Inhalt, Abstrakt und Umfang. Er liest das zu erschließende Dokument quer und erkennt dabei Häufungen von Schlüsselbegriffen. Er zerlegt nebenbei Komposita in ihre Bestandteile, reduziert Wortvarianten auf ihren Ursprung. Wesentliche Sinnträger erkennt er mühelos und Stoppwörter werden von ihm selbstverständlich eliminiert. Er erkennt ohne ins Detail gegangen zu sein, den inhaltlichen und thematischen Schwerpunkt des Dokuments. Dazu bedient er sich seines fachlich breit angelegten Hintergrundwissens, denn er kennt bereits eine Menge von Reizwörtern, die er mit einer einschlägigen Liste, vielleicht sogar einem Fachthesaurus oder einer Klassifikation eng verbinden kann.

Der Fachreferent muss nicht das Detailwissen des Autors haben, um die Hauptlinien eines Dokumentes grob zu verstehen.

Über Hilfsmittel, wie einen Thesaurus oder die Schlagwortnormdatei, kann er somit rasch über Stammformreduktion und/-oder Wortassoziationen zu Vorzugsdeskripto-

ren, die das Dokument hinreichend beschreiben. Dies sind nicht immer zwangsläufig Elemente die im Text vorkommen.

Als Statistiker, der er in dieser Tätigkeit ist, erkennt er Häufigkeiten von Begriffen und verwandte Begriffe und schätzt kontextabhängig deren Relevanz. Er kann aus der syntaktischen Struktur des Dokumentes, also der Position der Reizwörter, Semantik ableiten und die Relevanz des Vorkommens der Begriffe im Dokument bewerten und kann schließlich über Ähnlichkeitsverfahren und Normalisierung, des von ihm im Kopf erzeugten ‚Index' zur fest definierten Fachterminologie gelangen. Er verwendet generative und partitive Beziehungsattribute, um die dem Ganzen am nächsten kommenden Schlagworte oder Klassifikationsmerkmale zu wählen. Dazu bemüht er auch sein Kontextwissen um das Dokument, das sowohl zeitlich als auch inhaltlich und räumlich anzuwenden ist. Dieser Kontext eines Dokuments ist allerdings nicht statisch und auch nicht objektiv. Anhand der von ihm als wesentlich empfundenen Terme, gelingt ihm eine Abbildung bzw. Zuordnung zu einem in seiner Institution fest vorgegebenen Klassifikationssystem, das er bei langjähriger und häufiger Anwendung kennt.

Welche Instrumente benutzt der Fachreferent hier intuitiv bei der intellektuellen Zuordnung eines Dokuments zu einer vorhandenen Klassifikation.

- Er wendet linguistische und statistische Verfahren des IR (Information Retrival) an, wie z. B. die Stammformreduktion oder die Zerlegung in Komposita.
- Er bewertet Relevanz und Häufigkeit von Reizwörtern, durch statistische Methoden. Die Beziehungen der Begriffe werden hierarchisch abgebildet. Es kann sich dabei, je nach Kontext, um beliebige Beziehung handeln.
- Er versucht eine Abbildung der so normalisierten Worte, mit Hilfe semantischer Analyse auf einem definierten Wortschatz, der zur Erschließung genutzt werden soll, wie z.B. Schlagwortnormdatei, Klassifikationen...

In den folgenden Kapiteln wird der Versuch unternommen, das Szenario mit einer Reihe von Verfahren zu beschreiben, die für den Entwurf einer automatisierten Klassifikation genutzt werden sollen.

2 Eigenschaften von Fachsprachen

Die zu klassifizierenden Textdokumente aus dem Fachbereich Informatik, bedienen sich, wie jeder Fachtext, einer Fachsprache. Fachsprachen bauen, zur eindeutigen Kommunikation im Fachgebiet, auf einem eigenen System von Begriffen auf. Die Gesamtheit der Begriffe eines Fachgebietes wird als Fachterminologie bezeichnet.

Eine Terminologie ist nach Eugen Wüster *„…ein Begriffs- und Benennungssystem, das alle Fachausdrücke umfasst, die für dieses Fachgebiet allgemein üblich sind…"*[WÜST91]. Dabei entspricht ein Begriff der Bedeutung der Worte, wie wir sie verstehen und die zugehörige Benennung ist die Zeichenform. Zusammen bilden Begriff und Benennung einen Term, der aus einem oder mehreren Worten bestehen kann. Die Zuordnung von Begriff und Benennung ist nicht immer exakt. Die Beziehungen zwischen Termen und deren hierarchische Anordnung kann durch Definitionen exakt beschrieben werden. Auch die Problematik von Synonymen und Homonymen gehört in dieses Gebiet.

Eine epische Darstellungsform ist, im Gegensatz zu Texten schöngeistigen Inhalts, insbesondere in technischen Fachgebieten, nicht üblich, da das Anliegen eines Fachtextes ja darin besteht, einen Sachverhalt möglichst exakt zu beschreiben. Mehrdeutigkeiten und Synonyme sind in der Fachterminologie zwar ebenfalls vorhanden, jedoch nicht in dem Maße, wie in der Gemeinsprache, die nach DIN 2342 als *„Kernbereich der Sprache, an dem alle Teilnehmer einer Sprachgemeinschaft teilhaben"* definiert ist.

Für die Extraktion von Begriffen aus Fachtexten ist es notwendig, deren Unterschiede in den Eigenschaften gegenüber den Eigenschaften gemeinsprachlicher Texte zu berücksichtigen. Allerdings ist es für den einzelnen Begriff oft schwer abzugrenzen, wann dieser zur Fachterminologie gehört und wann es ein gemeinsprachlicher Begriff ist. Die Grenzen sind fließend und Begriffe können auch aus der Fachterminologie in den allgemeinen Sprachgebrauch hinüberwandern, bzw. in beiden vorkommen. Beispiele in der Informatik für einen Wechsel vom Fachterminus zum gemeinsprachlichen Begriff, sind vor allem im Bereich Internet und WWW zu finden.

Für die statistische Analyse von Fachtexten können bestimmte Eigenschaften ausgenutzt werden. Fachspezifische Termini sind statistisch anders verteilt, als gemeinsprachliche Worte. Fachterme werden nach bestimmten Mustern gebildet, deren Erkennung beim Auffinden dieser Terme ausgenutzt werden kann. Es werden oft Abkürzungen verwendet, die der Sprachökonomie dienen, aber trotzdem noch eindeutig sind. Komposita, die im deutschen Sprachschatz oft vorhanden sind und Mehrworttermen in anderen Sprachen entsprechen, haben also einen größeren Informationsgehalt. Mehrwortterme werden in Fachsprachen öfter als in gemeinsprachlichen Texten angewandt. Diese sind mit größerer Wahrscheinlichkeit fachlich relevant als Einwortterme.

Für die Bedeutung von Termen ist die statistische Betrachtung der Verteilung der Terme von Bedeutung. Dabei gilt das Zipfsche Gesetz, welches folgende Aussage beinhaltet: *„…Werden Wörter eines Textes in der Rangfolge ihrer Häufigkeit aufgelistet, ist die Häufigkeit eines Wortes umgekehrt proportional zu seiner Rangstelle...“* Das Produkt aus Rang und Häufigkeit ist also konstant. Das Zipfsche Gesetz hängt eng mit der Pareto Verteilung zusammen, die aussagt, dass eine kleine Anzahl von hoch bewerteten Elementen in einer Menge sehr viel zum Gesamtwert der Menge beitragen, wohingegen der überwiegende Teil der Elemente nur sehr wenig zum Gesamtwert beiträgt. Das Zipfsche Gesetz gilt sowohl für Fachtexte als auch für Texte allgemeiner Natur.

In [HOFF88] werden Eigenschaften von Fachtexten gegenüber allgemeinen Texten untersucht. Dabei wurde unter anderem herausgearbeitet, dass

- Stoppwörter in allen Texten gleichermaßen häufig auftreten,
- Terme, insbesondere Nomina, die in Fachtexten häufig auftreten, in der Gemeinsprache nur sehr selten auftreten.
- Die Mehrdeutigkeit von Termen in Fachtexten signifikant geringer ist.

Diese Eigenschaften sollen bei der Analyse eines Fachtextes ausgenutzt werden.

3 Textindizierung und Merkmalsextraktion

Um vom Dokument zur Klassifikation zu gelangen, muss zunächst eine Zerlegung des Textes erfolgen. Dies beginnt immer mit dem Erzeugen eines Index auf dem Dokument. Aufgrund der Komplexität des Gebiets des Information Retrieval soll hier nur kurz auf die möglichen Verfahren eingegangen werden, die für das Szenario genutzt werden können.

Indexierung des Textes

Das Ursprungsdokument (Metadaten und der Volltext) wird zunächst auf eine unstrukturierte Menge von Termen abgebildet.
Die Umwandlung erfolgt in mehreren Teilschritten:

- Elimination der Struktur. Syntaktische Strukturelemente, wie z.B. XML-Tags werden für die Termextraktion nicht berücksichtigt.
- Elimination von häufigen Termen, den Stoppwörtern. Dies sind typischerweise Wörter mit geringer Aussagekraft, die zur Indexierung nicht verwendet werden, da sie in fast allen Dokumenten auftreten. Die 200-300 häufigsten Wörter einer Kollektion von Dokumenten sind ebenfalls solche mit geringer Aussagekraft. Zum Beispiel ist der Begriff „Informatik“ nutzlos für die Indexierung von Dokumenten, die nur dem Fachgebiet der Informatik zugehörig sind. Durch die Elimination von Stoppwörtern wird der Speicheraufwand reduziert und das Retrieval effizienter.
- Der Text wird in Terme aufgebrochen. Mögliche Termformen sind dabei: Wörter, Phrasen oder so genannte N-Gramme. Hier geht es um die Ermittlung von Wörtern, Wortfragmenten und Wortfolgen, die als Terme zur Beschreibung des Textdokuments herangezogen werden. Meist werden Terme in Form von Wör-

tern und Phrasen (mit Stammformreduktion, eventuell Groß/Kleinschreibung, Fehlerkorrektur) verwendet. Ein anderer Ansatz ist es Wortfragmente, so genannte N-Gramme, für die Indizierung zu nutzen:

Bsp für N=3 - Trigramme	
Klasse	kla,las,ass,sse
Klassen	kla, las, ass, sse, sen
Klase	kla,las,ase

- Reduktion der Terme auf Stammform, Zerlegung der Komposita. Für die englische Sprache gibt es allgemeingültige Algorithmen, wie zum Beispiel den Porter Algorithmus, der dazu genutzt werden kann, die verschiedenen Terme mit gleicher Stammform in ein und denselben Term zu überführen. Wegen der starken Konjugation und Deklination können deutsche Wörter nicht automatisch in ihre Stammform gebracht werden. Deshalb werden dafür zusätzlich Thesauri verwendet, die für jede Stammform die möglichen Ableitungen enthalten. Zudem benutzt die deutsche Sprache Komposita, d.h. zusammengesetzte Wörter, welche in ihre Bestandteile aufgespaltet werden müssen. Bei der Verwendung von N-Grammen, kann auf eine Stammformreduktion und eine Silbentrennung verzichtet werden.

Als Ergebnis liegt nach der Indexierung eine Menge von Begriffen vor. Die Begriffe sind, je nach Verfahren, auf ihre Stammform reduziert, oder es existiert eine Menge von Silben und Wortbestandteilen.

Terminologie Extraktion

Zur Bestimmung der Relevanz eines Begriffes werden die Position und das Auftreten innerhalb des Dokuments und in Bezug zum allgemeinen Sprachschatz als ein wesentliches Merkmal hinzugezogen.

Eine Relevanzgrenze wird erreicht, wenn die Termhäufigkeit im Dokument einen bestimmten festzulegenden Erwartungswert X um einen Faktor Y überschreitet

[QUAS03]. Diese Relevanz, also wie groß ist die Bedeutung eines Begriffes innerhalb eines Textes, wird über verschiedene Verfahren ermittelt. Ein Verfahren, das sich sowohl statistischer als auch linguistischer Methoden bedient ist die Differenzanalyse.
Die im Kapitel 2 Eigenschaften von Fachsprachen angeführten Aussagen zu Eigenschaften von Fachtexten, werden im Verfahren der Differenzanalyse, das unter anderem dazu dienen kann, fachspezifische Terme zu identifizieren, ausgenutzt.

Differenzanalyse

In [HEYE04] wir die Differenzanalyse für den Einsatz in der Spracherkennung folgendermaßen beschrieben:
Den Ausgangspunkt für die Differenzanalyse bilden zwei Textmengen: Der zu analysierende Text, aus denen die diskriminierenden Terme extrahiert werden sollen und ein Referenzkorpus. Der zu analysierende Text kann ein Fachtext sein, der Referenzkorpus hingegen ist ein gemeinsprachlicher Textkorpus.
Es ist also eine Analyse über Korpora unterschiedlicher Basis möglich. Um die Signifikanz eines Terms zu bestimmen ist es notwendig, einen größeren Textkorpus als Vergleichskorpus zu besitzen. Damit lässt sich die Auftretenswahrscheinlichkeit eines Terms in Bezug auf den Gesamtkorpus mit der tatsächlichen Auftretenshäufigkeit im Analysetext vergleichen und als Relevanzinformation nutzen. Die Verwendung eines sehr großen Korpus ist notwendig, um hinreichende Sicherheit bei der relativen Häufigkeit zu erlangen.
Werden nun die Auftretenswahrscheinlichkeiten einzelner Wortformen bzw. Wortformenkombinationen für die beiden Textkorpora berechnet und zueinander in Beziehung gesetzt, dann lassen sich grundsätzlich folgende vier Klassen von Wortformen bilden:

- Wortformen, die im Referenzkorpus nicht vorkommen. Sofern sie im Analysekorpus nicht nur einmal gesehen werden (z.B. Tippfehler), handelt es sich bei diesen Wortformen mit hoher Wahrscheinlichkeit um Fachterme des Themengebiets, über den im Fachtext berichtet wird.
- Wortformen, die im Analysekorpus relativ häufiger vorkommen als im Referenzkorpus. Bei diesen Wortformen handelt es sich mit einer gewissen Wahr-

scheinlichkeit um Fachterme. Für die Identifizierung dieser Wortformen muss ein Schwellwert definiert werden, z. B. eine Mindestdifferenz der Häufigkeitsklassen (vgl. unten) im Analyse- bzw. Referenzkorpus.

- Wortformen, die in beiden Textkorpora relativ gleich häufig vorkommen. Dabei handelt es sich meist um Stoppwörter (Wortformen aus den geschlossenen Wortklassen Artikel, Konjunktion und Präposition) oder um allgemeine Begriffe. Diese Wortformen vermitteln keine themenspezifischen Inhalte des Fachtextes.
- Wortformen, die im Fachtext seltener als im gemeinsprachlichen Textkorpusvorkommen sind im Allgemeinen keine diskriminierenden Fachterme.

Für die Anwendungen der Differenzanalyse zur Terminologie Extraktion sind die Wortformen aus den Punkten 1 und 2 wesentlich.

Hinsichtlich der Relevanz der Worte für einen Fachtext sind die 100 häufigsten Worte eines Fachtextes genauso wenig inhaltstragend, wie die 100 häufigsten Worte eines Allgemeintextes. Das heißt man kann hierbei für die häufigsten Worte die gleichen Eigenschaften annehmen, dass sie also unspezifisch sind. Die Begriffe, die im Fachtext signifikant häufiger als im allgemeinen Wortschatz vorkommen, sind für die Relevanz von großer Bedeutung.

Dieses Auftreten wird als Häufigkeitsklasse (HKL) einer Wortform bezeichnet und lässt sich mathematisch wie folgt ausdrücken:

$$HKL(w) = \text{ganzer Anteil } (\log 2\ (|'der'|/|w|))$$

mit $|w|$ als Anzahl der Vorkommen der Wortform *w*.

Das bedeutet, dass die Wortform '*der*' als häufigste Wortform im Korpus etwa $2^{HKL(w)}$ mal so oft auftritt, wie die Wortform *w*. Nach dem Zipfschen Gesetz erhält man die gleiche Einteilung der Wortformen in Häufigkeitsklassen, wenn man in der häufigkeitssortierten Liste der Wortformen alle Wortformen von Rangplatz 2^{i+1} bis Rangplatz 2^i in die Häufigkeitsklasse *i* einordnet.

Die Analyse lässt sich durch weitere Filtertechniken noch verbessern. Zum Beispiel ist es sinnvoll nur alle Nomina zu betrachten, oder die Häufigkeit des Auftretens eines

Terms mit einem Mindestwert zu belegen. So können z.B. Rechtschreibfehler ausgeschlossen werden. Sinnvoll ist es eine statistische Prüfgröße zu ermitteln, die ein Ergebnis der Differenzanalyse bewertet. Die differenzierte Betrachtung von Einwort- und Mehrworttermen wäre sinnvoll.

Nach Anwendung aller Methoden erhält man eine Liste von Worten, die die relevante Terminologie des Textes darstellen. Die entstanden Liste entspricht nach [WITS05] in etwa der Beschlagwortung eines Textes mit freien Schlagworten, die für die Beschreibung des Inhalts relevant sind.

4 Strukturierende Instrumente

Ein weiterer wesentlicher Schritt ist nun die Abbildung dieser Liste von Termen auf eine Klassifikation bzw. ein System, das zur inhaltlichen Erschließung des Dokuments in Bibliotheken genutzt werden kann. Idealerweise nutzt man dazu als Ausgangsbasis einen Fachthesaurus, da dieser eine Klassifikation impliziert.
Hierin besteht ein wesentlicher Unterschied zu anderen Verfahren, die anhand einer Termextraktion aus den vorliegenden Fachtexten eine Klassifikation, einen Thesaurus oder ein Begriffsnetz versuchen zu erzeugen. Als Online- Beispiel ist hier das Projekt Dandelon http://www.dandelon.com zu nennen.
Da in diesem Kontext oft Begriffe wie Thesaurus, Ontologie, Klassifikation, Topic Map fallen und teilweise synonym verwendet werden, werden im folgenden Abschnitt die Instrumente Thesaurus und Klassifikation sowie Begriffe wie Ontologie und Topic Map kurz vorgestellt und anschließend bezüglich ihrer Eigenschaften vergleichen.

In Thesauri und Klassifikationen ist die Unterscheidung von Begriffen und deren natürlich sprachlicher Benennung sehr wichtig. Oft werden Objekte einer Klassifikation, einer Topic Map oder eines Thesaurus mit einer Art Thema oder Aussage beschrieben,

dessen einzelne Begriffe oder Wortbestandteile überhaupt nicht den Inhalt widerspiegeln. Die gegenseitige Zuordnung von Begriffen und Benennungen ist jedoch notwendig, um Fachtexte definierten Begriffen zuordnen zu können.

Klassifikationen

Als eine Klasse bezeichnet man einerseits eine Menge von Objekten, die aufgrund gemeinsamer Merkmale, in Abgrenzung zu anderen Objekten zu einer Gruppe zusammengefasst werden, andererseits aber auch das Merkmal oder die Eigenschaft selbst [WIKI06]. Im bibliothekarischen Sinn sind diese Objekte die Begriffe bzw. Vorzugsdeskriptoren.
Klassifikationen dienen dem Ordnen von Objekten und der vereinfachten Abbildung in ein Ordnungsmodell, da im Normalfall die Anzahl der zu klassifizierenden Objekte größer ist als die Anzahl der Klassen. In einem Klassifikationssystem gibt es mindestens zwei Arten von Klassenbeziehungen, die syntaktische Beziehung und die hierarchische Beziehung.
Hierarchische Beziehungen dienen dazu generative, partitive oder assoziative Ordnungen oder aber auch beliebig attributierte Beziehungen der Klassen untereinander auszudrücken. Bei einer Hierarchie mit Einfachvererbung, besitzt jede Klasse nur eine Oberklasse, so dass die gesamte Klassifikation eine Baumstruktur besitzt. Die Klassen in einer monohierarchischen Struktur sind meist generativ oder partitiv. Bei der Hierarchie mit Mehrfachvererbung, auch Polyhierarchie genannt, kann eine Klasse mehreren Oberklassen untergeordnet werden. Klassifikationen, die polyhierarchisch aufgebaut sind können wesentlich mehr Aspekte von Ordnungen darstellen und lassen damit fast beliebige Betrachtungsweisen zu. Die Vielfältigkeit polyhierarchischer Klassifikationen wird vor allem bei deren computerbasierten Nutzung sichtbar, da grundsätzlich beliebig Merkmale miteinander kombiniert werden können und damit eine Sicht für jeden beliebigen Aspekt, die die Klassifikation beinhaltet auf die klassifizierten Dokumente möglich ist.

Syntaktische Beziehungen von Klassen einer Klassifikation werden nach [BUCH89] in Einfachklassen mit Elementar- und differenzierten Klassen und komplexen Klassen mit Verbund- und interaktiven Klassen eingeteilt. Elementarklassen beschreiben ein Merkmal, differenzierte Klassen beschreiben mehrere Merkmale bezüglich eines Objektes. Komplexe Klassen beziehen sich auf die Beschreibung eines oder mehrerer Merkmale verschiedener Objekte.
Je nach Art der Erzeugung der Klassifikation unterscheidet man zwischen prä- und postkombinierte Klassifikationen. Präkombinierte Klassifikationen sind Klassifikationen in denen bereits alle Klassen von vornherein festgelegt sind. Da sich das Wissen und dessen Ordnungssystem darauf ständig verändern und erweitert, müssen diese Klassifikationssysteme laufend aktualisiert werden und sind dadurch nicht sehr flexibel. *„Ein Klassifikationssystem, das bereits alle Relationen in der Welt des Wissens, alle Phänomene also, berücksichtigt und dabei monohierarchisch einordnet, ist nicht möglich.“* [RANG69]
Postkombinierte Klassifikationen verzichten daher von vornherein darauf, Themen in vorgeprägten Klassen auszudrücken: Sie enthalten nur eine beschränkte Anzahl von vorgegebenen Klassen (Basisklassen), die dann miteinander kombiniert werden können. Mit so einem offenen Klassifikationssystem kann prinzipiell jeder Aspekt des klassifizierten Gegenstands gleichberechtigt berücksichtigt werden. Für eine Aufstellungssystematik sind diese Klassifikationen allerdings ungeeignet, bzw. die Mächtigkeit der Klassifikation spiegelt sich in der Aufstellung nicht wider.

Eine Klassifikation besitzt als Klassen im Allgemeinen nur Vorzugsdeskriptoren. Die ausschließliche Nutzung nur der Vorzugsdeskriptoren hat für die praktische Nutzung jedoch Nachteile, wenn von einer anderen Benennung als der der genannten Klasse ausgegangen wird. Aus diesem Grunde werden bei Klassifikationen teilweise die Klassen mit weiteren beschreibenden Benennungen angereichert. Durch eine Anreicherung der Klassen mit weiteren Deskriptoren und einer mehrdimensionalen Hierarchie der Klassen untereinander, nimmt die Klassifikation eine Entwicklung zu einem Thesaurus. Durch Hinzufügen eines nicht sprachlichen Schlüssels (Notation), ist es möglich, Klassifikationen letztendlich sprachunabhängig zu verwenden, so dass Übersetzungen prin-

zipiell nur für das bessere Verständnis notwendig sind, da die Zuordnung über den Schlüssel erfolgt.
Mit einer Fachklassifikation soll ein bestimmtes Fach oder Themengebiet hinreichend strukturiert werden. Die Klassifikation hat dabei keinen Anspruch auf Vollständigkeit.
Neben den intellektuell erstellten Klassifikationen, gibt es auch eine Menge automatisch erzeugter Begriffssysteme. Für bibliothekarische Anwendungen werden zumindest in Deutschland derzeit nur intellektuell erstellte Klassifikationen genutzt, da die genaue Zuordnung und das damit verbundene kontrollierte Vokabular für die Systematisierung wesentlich ist.

Thesaurus

„Ein Thesaurus im Bereich der Dokumentation ist eine Sammlung (Menge) von Begriffsbenennungen und/oder zusätzlicher Wörter der natürlichen Sprache und/oder sonstiger Zeichen mit Darstellungen von Beziehungen zwischen diesen Elementen, sofern noch folgende Kriterien erfüllt sind: Die Sammlung enthält einen merklichen Anteil von Nicht-Vorzugsbenennungen und/oder von nicht als Deskriptoren benutzten Vorzugsbenennungen. Eine terminologische Kontrolle wird angestrebt.“ [SOER69]
Unterschiedliche Schreibweisen (Photo/Foto), Synonyme bzw. als gleichbedeutend behandelte Quasi-Synonyme, Abkürzungen, Übersetzungen etc. werden durch Äquivalenzrelationen miteinander in Beziehung gesetzt. Begriffe werden außerdem durch Assoziationsrelationen und hierarchische Relationen vernetzt. In der Definition von Umstätter [UMST89] wird vor allem die Polyhierarchie eines Thesaurus als dessen wesentliches Merkmal gegenüber Klassifikationen hervorgehoben. Für meinen Entwurf sind beide Merkmale, die hohe Anzahl weiterer Deskriptoren und die Möglichkeiten, die eine Polyhierarchie bietet, wesentlich.

Der Thesaurus kann als geeignetes Hilfsmittel zur Sacherschließung und zum Auffinden von Dokumenten genutzt werden. Dabei dienen Relationen zwischen den einzelnen Begriffen zum Auffinden, bei der Vergabe von Schlagworten und bei der Recherche.

Die Thesaurusnormen DIN 1463-1 bzw. das internationale Äquivalent ISO 2788 sehen mindestens folgende Relationsarten und dazugehörige Abkürzungen vor:

Kürzel und Bezeichnung	
DIN 1463-1	**ISO 2788**
BF - Benutzt für	UF - Used for
BS - Benutze Synonym	USE/SYN Use synonym
OB - Oberbegriff	BT - Broader term
UB - Unterbegriff	NT - Narrower term
VB - Verwandter Begriff	RT - Related term
SB – Spitzenbegriff	TT - Top term

Die häufigsten Relationen in einem Thesaurus sind Äquivalenz-, Assoziations- und hierarchische Relationen.
Die Gesamtheit, der durch Deskriptoren repräsentierten Begriffe, die ein Thesaurus enthält und die zwischen ihnen bestehenden Beziehungen bilden die klassifikatorische Struktur des Thesaurus, die eine Obermenge des Klassifikationssystems sein kann.
Im Gegensatz zu einer monohierarchischen Tabelle kann der Thesaurus eine polyhierarchische Struktur besitzen, dass heißt ein Unterbegriff kann mehrere Oberbegriffe haben. Prinzipiell kann mit der polyhierarchischen Ordnung jeder beliebige Aspekt als Relation genutzt werden. Dadurch können beliebige Zusammenhänge besser dargestellt werden. Von jedem beliebigen Punkt des Benennungssystems könnte der Informationsgehalt, der den Begriffen zugeordnet ist, unter ganz verschiedenen Aspekten betrachtet werden. Dadurch können Zusammenhänge wesentlich realitätsnaher dargestellt werden als mit einem monohierarchischen System.
Für die verschiedenen Fachgebiete existieren einzelne Fachthesauri, wie z. B. in der Medizin der MeSH Thesaurus: **MeSH** steht für **Me**dical **S**ubject **H**eadings. Hierbei han-

delt es sich um einen Thesaurus, den die National Library of Medicine (NLM), USA erstellt und fortlaufend pflegt.
Für das Fachgebiet der Informatik existiert zurzeit kein publizierter Fachthesaurus.

Topic Maps

Eine Topic Map ist ein abstraktes Modell zur Formulierung von Wissensstrukturen. Ihren Ursprung haben Topic Maps in Klassifikationen und Thesauri.

Zur formalen Beschreibung des Modells gibt es folgende Bestandteile:

- Topics (Names, Scopes)
- Assoziationen
- Occurances

Topics sind dabei Themen, oder auch Subjekte die in der Regel fest definiert sind. Assoziationen stellen die Verknüpfungen zwischen den Topics dar. Diese Assoziationen stellen nur unbenannte Beziehungen dar. Als Occurances bzw. Instanzen werden die dahinter liegenden Dokumente, die dem Topic zugeordnet sind, bezeichnet.
Zu jedem Topic kann ein bestimmter Namensraum (Scope) bestimmt werden, indem die Begriffe gelten. Dadurch kann durch einfaches Wechseln des Scope, das z.B. die Sprache sein kann, eine andere Darstellung der Topic Map erreicht werden.
Mit Hilfe des XML Topic Map Standards XTM können Wissensstrukturen und Assoziationen in Informationsbeständen auf standardisierte Weise beschrieben werden.
Topic Maps können je nach Zweck, sowohl intellektuell erzeugt werden, als auch automatisch aus den zugrunde liegenden Instanzen generiert werden. Automatisch erzeugte Topic Maps stellen einen ‚begrenzten Auszug der Realität' dar, und sind meist polyhierarchisch. Die Beziehungen der extrahierten Begriffe entstehen vor allem aus der sta-

tistischen Nähe und der Position zueinander und innerhalb der zugrunde liegenden Datenquellen. Aufbauend auf eine Topic Map kann eine Ontologie modelliert werden.

Ontologien

Unter Ontologie versteht man im Bereich der Informatik und der Repräsentation von Wissen ein „formales Modell eines Wissensraumes“, das durch Konzepte und Relationen beschrieben wird. Eine Ontologie kann auf einer Topic Map, die als Wissensraum dient, aufsetzen und somit zur Beschreibung der Zusammenhänge dienen. Die Konzepte dienen dabei der formalen Beschreibung der Daten, die Relationen der Beschreibung des Zusammenhangs. Die Relationen können beliebig attributiert werden. Dadurch kann je nach Aspekt ein ‚Begriffsnetz' bezüglich bestimmter Relationen aus sehr verschiedenen Ansichten betrachtet werden.
Ontologien können nicht automatisch generiert werden, da die Relationen intellektuell festgelegt werden müssen, da diese bei den zugrunde liegenden Topic Maps unbenannt sind. Die Beziehungen können derzeit noch nicht befriedigend aus den Datenquellen generiert werden.

Vergleich von Strukturierungssystemen

In der Praxis finden vor allem aus dem Dokumentenbestand automatisch erzeugte Begriffsnetze Anwendung, die als einfache Topic Maps bezeichnet werden können. Die Definition darüber liegender Ontologien für die begrenzte Darstellung eines Themas und der Beziehungen der Begriffe untereinander, erfolgt in der Regel durch umfangreiche kreative Tätigkeiten, die vergleichbar sind mit der Erstellung einer Fachklassifkation. Automatisch erzeugte Begriffsnetze dienen der Strukturierung großer Mengen von Textdokumenten und als Hilfsmittel für eine intelligentere Suche auf diesem Bestand.

Ein Merkmal ist hier, das die Begriffsnetze für in den Dokumenten enthaltene Informationen entstehen, aber keine Aussagen zu nicht vorhandenen Begriffen, da die Begriffsnetze nur vom Inhalt der Dokumente gespeist wurden. Die Beziehungen sind unbenannt.

Für den Entwurf des Systems zur automatischen Klassifikation, ist es vor allem wichtig, neben den Vorzugsdeskriptoren, genügend weitere Deskriptoren zu besitzen, um so eine möglichst gute Abbildung verwandter Terme zu erreichen.

Da der Thesaurus als Obermenge der vorgestellten Systeme angesehen werden kann, soll kurz ein Vergleich der Eigenschaften bezüglich der Thesaurusmerkmale dargestellt werden.

Merkmalsvergleich von Begriffssystemen			
Thesaurus	**Topic Map**	**Ontologie**	**Klassifikation**
Used for	Topic	Vorzugsdeskriptor	Klasse
Use synonym	-	Deskriptor	[Terms]
Broader term	-	Benannte Relationen	Vaterknoten
Narrower term	-	Benannte Relationen	Kindknoten
Related term	Association	Benannte Relationen	[Is related to]
Top term	-	Root	Top

Ontologie und Klassifikationen, besitzen bezüglich der Thesaurusdefinition ähnliche Eigenschaften. Thesauri sind meist noch mit einer Definition des Vorzugsdeskriptors angereichert. Dies ist bei Ontologien und Klassifikationen nicht der Fall. Ontologien besitzen bezüglich ihrer Relationen einen weiteren Freiheitsgrad, da die Relationen prinzipiell beliebig attributiert sein können. Klassifikationen sind zwar in der Praxis meist monohierarchisch, die Möglichkeit der Polyhierarchie ist aber in einigen Klassifikationen durch weitere Attribute, wie [is related to] vorgesehen. Topic Maps haben gegenüber Thesauri nur wenig adäquate Eigenschaften. Ihre Beziehungen sind im Allgemeinen unbenannt. Das hat vor allem seinen Ursprung in der Erzeugung. Thesaurus, Klassifikation und Ontologie sind Produkte kreativer Arbeit, Topic Maps können automatisch erzeugt werden. Ihre Instanzen, also die für das Topic relevanten Texte sind bereits im Modell vorgesehen, während Klassifikationen, Thesaurus und Ontologie zunächst ohne dahinter liegende Datenquellen erzeugt werden. Hier erfolgt die Zuordnung nach dem Entwurf.

Prinzipiell ist es natürlich vom Anwendungsszenario abhängig, welche der genannten Instrumente zum Einsatz kommen. Der Aufwand eine Topic Map zu erzeugen ist wohl am geringsten, es wird allerdings auch kein definiertes Vokabular verwendet und bildet nur den aktuellen Informationsgehalt ab, nicht jedoch ‚Informationslücken'.

5 Auswahl einer fachspezifische Klassifikation

Für die Auswahl der Klassifikation, ist vor allem die Verbreitung und Nutzbarkeit der Klassifikation für das Anwendungsszenario wesentlich. Für eine automatisierte Klassifikation von Online-Dokumenten ist eine Aufstellungssystematik irrelevant. Polyhierarchisch strukturierte Klassifikationen können vor allem in Online Anwendungen optimal ausgenutzt werden.

Im Fachgebiet der Computerwissenschaften ist das Computing Reviews Classification System (CRCS) der ACM (Association for Computing Machinery) – auch CR Classification genannt - heute das am meisten angewandte Klassifikationssystem und bildet damit eine Quasistandard [ACM 98] in diesem Bereich.

Die CR Classification wurde von der Association for Computing Machinery ACM erstellt und eine erste Version 1964 publiziert. Die Klassifikation wurde und wird sukzessive weiterentwickelt. Die aktuelle Version basiert auf dem Entwurf von 1998. Die CR Classification ist frei verfügbar und liegt als Hypertext, als ASCII Text und in XML-Notation vor.

In der XML Notation, wird der Aufbau am anschaulichsten dargestellt. Die CR Classification besitzt die Elemente <node>, <is composed by> und <is related to>, die genügen um die Beziehungen und Benennungen in einem hierarchischen System abzubilden.

Node Element der CR Classification in XML Notation:

```
<node ... >
 <node id={CR_Descriptor} label={CR_Label}>
  <is composed by>
   {non descriptor Text}
  </is composed by>
  <is related to node={CR-Descriptor} />
 </node>
</node>
```

Sie stellt ein hierarchisches strukturiertes Begriffssystem dar. Die Vorzugsdeskriptoren sind eindeutig und verweisen auch durch ihre Schlüsselsyntax auf den nächst höher liegenden Vorzugsdeskriptor. Die Struktur ist baumartig aufgebaut und hat eine Tiefe von 3 Ebenen. In der 4. Ebene wird der Vorzugsdeskriptor des Knotens durch weitere Elemente beschrieben. Weiterhin gibt es Verweise auf die Zugehörigkeit eines Vorzugsdeskriptors zu weiteren Vaterelementen, als seinem direkten Vaterknoten. Dadurch ermöglichen die <is related to> Elemente eine Polyhierarchie. <is composed by> Elemente dienen der Zuordnung von Synonymen und weiteren Deskriptoren.

Die CR Classification liegt ausschließlich in englischer Sprache vor, wobei durch die Nutzung der eindeutigen Identifikatoren, prinzipiell keine Übersetzung notwendig ist, sie diente lediglich dem leichteren Verständnis.

Die XML Variante ist für ein automatisches Klassifikationsverfahren am besten geeignet, da ja die Semantik der Elemente durch die XML-Notation ausgedrückt wird und die Nutzung technologisch am einfachsten ist.

Ausschnitt der CR Classification XML Notation für ein Node Element:

```
<node id="H." label="Information Systems">
 <isComposedBy>
  <node id="H.0" label="GENERAL"/>
  <node id="H.1" label="MODELS AND PRINCIPLES">
   <isComposedBy>
    <node id="H.1.0" label="General"/>
    <node id="H.1.1" label="Systems and Information Theory">
     <isRelatedTo>
       <node id="E.4"/>
     </isRelatedTo>
     <isComposedBy>
       <node label="General systems theory"/>
       <node label="Information theory"/>
       <node label="Value of information"/>
     </isComposedBy>
    </node>
    ...
   </isComposedBy>
  </node>
  ...
 </isComposedBy>
</node>
```

6 Fachtermini und Klassifikation

Die CR Classification als solche ist nicht ausreichend für eine automatisierbare Zuordnung von Fachtexten zur Klassifikation, wie sie im Entwurf vorgesehen ist. Aus diesem Grunde wird die Klassifikation soweit mit Fachtermen angereichert, dass sie sich sukzessive zur Fachterminologie entwickelt, deren hierarchische Beziehungen durch die Klassifikationsstruktur ausgedrückt wird. Die Fachterminologie wird quasi in das Schema der Klassifikation ‚gedrückt', bzw. es wird eine Sicht auf die Fachterminologie über die Klassifikation erzeugt. Das ist insofern sinnvoll, da ja die Zuordnung der Fachtexte

zur Klassifikation erfolgen soll. Das Beziehungsgeflecht der Fachterminologie der Informatik ist natürlich wesentlich komplexer strukturiert. Die Erzeugung dieser hierarchischen Darstellung der Fachterminologie ist ein wesentlicher Schwerpunkt für dieses Verfahren zu automatischen Klassifikation. Es ist die Basis für die Klassifikation der Dokumente.
Bei einer Mindestanreicherung der Klassifikation mit Termen aus Fachtexten könnte zumindest eine semi-automatische Zuordnung von Fachtexten zur Klassifikation erreicht werden.
Dazu gibt es verschiedene Varianten:

Anreicherung der Fachtextdatenbasis:
Fachtexte, die bereits klassifiziert sind, könnten als Instanzen der jeweiligen Klasse hinterlegt werden um so eine bessere Datenbasis zu schaffen. Die Begriffe von Fachtexten, die bereits intellektuell klassifiziert sind, kann man als Deskriptoren den jeweiligen Klassen hinzufügen, um diese anzureichern.
Nach [SCHE05] liegt diese Menge bei ca. 15-20 Dokumenten pro Klasse, um eine relativ gute Trefferquote bei einer automatischen Zuordnung zu erreichen. Diese Werte beziehen sich auf ein Verfahren der Klassifizierung von Pressetexten. Wenn man diese Werte auch für das hier zu entwerfende Verfahren annehmen kann, dann ist eine Vorweganreicherung der Klassifikation mit Fachtermini eventuell nicht notwendig, da der Mindestwert durch die Anwendung des Verfahrens ja relativ schnell erreicht wird.

Verbesserung der Extraktion und sukzessive Anreicherung der Datenbasis:
Mit jeder Klassifizierung, die erfolgt, werden die neuen, noch nicht zugeordneten Terme des Fachtextes als weitere Deskriptoren verwendet und das klassifizierte Dokument als Instanz zugeordnet. Dadurch wird das System lernend, da sich mit jedem klassifizierten Dokument die Basis der Fachtermini vergrößert und so die Automatisierung verbessert werden kann. Die gefundenen Klassen sollen jedoch auch immer als Vorschlag im Prozess des Klassifizierens dienen, um Fehler korrigieren zu können. In meinem Entwurf möchte ich versuche die verschiedenen Möglichkeiten zu berücksichtigen.

7 Datenquellen zur Anreicherung

Welche Datenquellen könnten zur Anreicherung herangezogen werden? Datenquellen müssen bestimmten Vorraussetzungen genügen, um als Anreicherung für eine Klassifikation herangezogen werden zu können. Ein wichtiger Punkt ist die Seriosität der Wissensquellen. Ein weiter Punkt, die Möglichkeit der freien Nutzung. Aus praktischer Sicht, sind auch die Integrationsmöglichkeiten in bestehenden Systemen, also die technologische Architektur der Wissensquelle, wichtig.
Große Datenquellen sowohl im populärwissenschaftlichen Bereich als auch im gewissen Maße im Fachbereich könnten offen zugängliche Wissensquellen sein. Ein bekanntes Beispiel dafür ist die Wikipedia.
Ein weiteres im Bibliotheksumfeld gebräuchliches Werkzeug ist die Schlagwortnormdatei, ein Instrument das nach den Regeln für den Schlagwortkatalog [RSWK98] angewandt wird.

Regeln für den Schlagwortkatalog

Das in Deutschland gebräuchlichste Werkzeug zur verbalen Sacherschließung sind die Regeln für den Schlagwortkatalog mit der zugehörigen Schlagwortnormdatei SWD. Die SWD wird hinsichtlich ihrer Eignung untersucht, ob sie den Anforderungen an einen Thesaurus bzw. an eine Klassifikation entspricht, bzw. ob sie als Hilfsmittel für die Automatische Klassifikation genutzt werden kann.

Auszug aus den Grundregeln:

1. *Die RSWK regeln die Inhaltserschließung von Bibliotheksbeständen durch die Schlagwortkatalogisierung. Ausgangspunkt ist die Praxis Der Deutschen Bibliothek und der Bibliotheksverbünde, die ihre Bestände unter Nutzung der Schlagwortnormdatei (SWD) erschließen. ...*
2. Die Schlagwortkatalogisierung kann nicht alle Beziehungen zwischen Begriffen und Gegenständen darstellen. Daher sollten die Dokumente daneben auch klassifiziert werden.
3. Die Schlagwortkatalogisierung kann durch maschinelle Indexierung ergänzt werden, insbesondere bei speziellen Gattungen von Dokumenten, wie retrokonvertierten Altbeständen, Zeitschriftenaufsätzen oder elektronischen Publikationen. Hier hat die maschinelle Indexierung in erster Linie die Funktion, Sucheinstiege zu vermehren, falls der Aufwand für eine intellektuelle Schlagwortvergabe unvertretbar hoch erscheint. Soweit durch maschinelles Indexieren keine rasch les- und interpretierbaren Inhaltsbeschreibungen im Sinn von § 13 erzeugt werden, ist eine Trennung beider Datenschichten für das Retrieval sinnvoll.
4. Die RSWK haben Bezüge zu den „Regeln für die Alphabetische Katalogisierung" (RAK). Die Erfassung der Daten nach beiden Regelwerken wird aufeinander abgestimmt.
5. Die RSWK berücksichtigen vorrangig die Bedürfnisse von Online-Katalogen…[RSWK98]

Eignung der SWD für eine automatisierte Klassifikation

Das wichtigste Instrument für die Arbeit des Fachreferenten mit der RSWK ist die SWD. Sie ist nach den oben genannten Regeln aufgebaut und soll danach verwendet werden. Seit der Schaffung der RSWK und dem damit verbundenen Aufbau, der Wartung, Aktualisierung und Pflege der SWD sind viele Änderungen vorgenommen worden, die vor allem versuchen, den Anforderungen und Bedürfnissen von Online-Katalogen gerecht zu werden.
Für die weiteren Betrachtungen soll nur die Seite der Sacherschließung und vor allem auch mit Blick auf Online-Dokumente betrachtet werden.
Die SWD ist kein Thesaurus und erst recht kein Fachthesaurus, da sie fachübergreifenden Charakter hat. Sie ist aus kontrolliertem Vokabular, den Begriffen, aufgebaut und

dient der Schaffung von Strukturen und Ordnungen auf den zugeordneten Dokumenten. Ihr fehlen neben den Beziehungen, und dem Beziehungsaspekt auch oft die Definitionen der Begriffe. Es existieren nur syntaktische Beziehung, bzw. Abstraktionsbeziehungen wie <Unterbegriff>, <Oberbegriff>, <Verweisung>, <Homonym>, <verwandter Begriff>, <verwendet für> zwischen den Deskriptoren.

Assoziative- oder Äquivalenzbeziehungen oder noch besser beliebig attributierte Beziehungen existieren nicht, bzw. werden assoziative Beziehungstypen nur sehr selten verwendet. Vor allem hinsichtlich dieser fehlenden Strukturinformationen ist hier ein großes Defizit.

Durch die Bildung von Schlagwortketten kann zwar eine Beziehung dargestellt werden, wobei diese jedoch im Interpretationsspielraum des Bibliothekars, und letztendlich des Anwenders liegt. Diese wird aber, im Gegensatz zu einer hierarchisch aspektorientierten Beziehung eines Thesaurus, gerade beim Online-Retrieval wieder zerschlagen und damit bedeutungslos. Einfache Verknüpfungen mit booleschen Operatoren können den semantischen Wert einer Beziehung nicht ersetzen.

Es werden der Pragmatik zuliebe beliebige Komposita gebildet, die auch mit mehreren Einzelschlagworten syntaktisch verknüpft den gleichen Sachverhalt darstellen können, z.B.: ‚Hochgeschwindigkeitsaerodynamik'.

Die Anwendung solcher Komposita auf Nutzerseite ist jedoch relativ unwahrscheinlich.

Der Informationsgehalt der SWD basiert nur auf dem Informationsgehalt der vorhandenen und eingearbeiteten Dokumente. Neueintragungen werden nur von einer relativen kleinen Gruppe von Nutzern vorgenommen.

Linguistische Verfahren des IR sind nicht vorhanden bzw. müssten erst von der jeweiligen Einrichtung implementiert werden, obwohl diese ja für alle Anwender gleich wären. Es existiert auch bezüglich der Recherchemöglichkeiten z.B. keine Wortstammreduktion, oder die Anwendung von Wortstammzerlegungen wie z.B. Trigrammen, die es ermöglicht Ähnlichkeitssuchen auf der SWD auszuführen und dem Nutzer bei der Suche im OPAC sicher weiterhelfen würden.

Das Anwendung des so genannten ‚engen Schlagworts' führt nur sehr bedingt zu einer guten sachlichen Erschließung. Je nach Umfang der Dokumentmenge müssten Schlag-

worte spezifischer gesetzt werden, d.h. …*die Zahl der Dokumente und die Zahl der Erschließungsbegriffe müssen in einem ausgewogenen Verhältnis stehen*. [UMST90]

Es ist sehr personalintensiv, die SWD unter Berücksichtigung des Regelwerks ständig zu aktualisieren, zu erweitern und zu pflegen und den aktuellen Ansprüchen, den sie bei der Schlagwortvergabe aktueller Literatur genügen muss, nachzukommen. Bei der Arbeit als Fachreferent, insbesondere bei der Verschlagwortung von aktueller Literatur in technischen Fachgebieten, werden regelmäßig Grenzen erreicht, da neu entstehende oder entstandene Wissensgebiete in der SWD unzureichend abgebildet sind. Das Einpflegen neuer Schlagworte unterliegt einer langwierigen Prozedur, die am Ende auch die Ablehnung des Schlagwortes mit sich bringen kann.

Im Zusammenhang mit der automatisierten Anwendung der SWD auf den zum Dokument gehörigen genormten Index, gibt es Probleme, da die SWD zwar online über die Katalogisierungsverbünde verfügbar ist, es aber keine Schnittstellen für automatische Anfragen gibt. Eine Integration in eigene Anwendungen ist, zumindest nicht über eine offene, definierte Schnittstelle möglich. Die bei einer Abfrage über einen Browser gelieferten Ergebnisse sind bezüglich ihrer Semantik ebenfalls schwer auswertbar. Es wird nur eine ungenügende Auszeichnung vorgenommen. Auch Verweise sind soweit vorhanden nur in reiner Textform dargestellt. Hier wäre es z.B. sehr hilfreich wenn Abstraktionsbeziehungen über ein XML Schema definiert und in XML Notation ausgezeichnet wären, so dass gezielt im Ergebnis, danach mit entsprechenden Parsern gesucht werden kann. Eine solche API, die das Ergebnis einer Anfrage als XML liefert wäre ein großer Vorteil hinsichtlich der Verwendung in eigenen Applikationen. Dann könnten die Beziehungen zwischen den Begriffen, die ja vorhanden sind, ausgenutzt werden. Eine Navigation über die SWD hin zum relevanten Dokument wäre dann relativ einfach umsetzbar. Insofern ist es unbefriedigend, dass diese große Datenbasis definierten Vokabulars derzeit nur so schlecht zu integrieren ist.

Es gibt also eine Reihe von Merkmalen und Regelungen bezüglich der SWD, die für dieses Anwendungsszenario nachteilig sind. Insgesamt ist die SWD in ihrer derzeitigen Art als Hilfsmittel für eine automatische Klassifikation von Online-Dokumenten nicht geeignet.

Wikipedia

Im Informations-Zeitalter haben sich neben dem Instrument RSWK und SWD andere, für die Inhaltserschließung, gerade von Online-Dokumenten durchaus interessante Werkzeuge entwickelt, oder sind in der Entwicklung und auch schon in Teilen etabliert. Inwieweit diese Werkzeuge als Hilfsmittel für eine automatische Klassifikation herangezogen zu werden könnten, soll kurz untersucht werden.

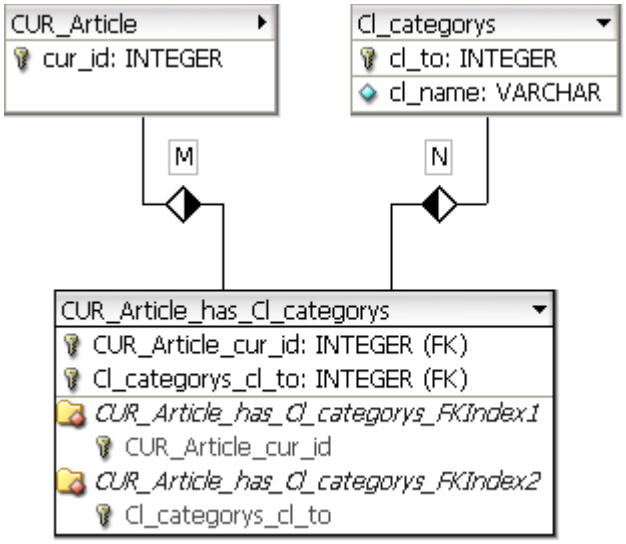
Das Wiki ist eine freie Software unter GNU public License. Eine Anwendung davon, die ebenfalls frei zur Verfügung steht ist Wikipedia und ihre Nachbarn wie z.B. Wikidictionary. Hier soll Wikipedia nur daraufhin näher untersucht werden ob es den Anforderungen an einen Thesaurus bzw. den Anforderungen an eine Klassifikation entspricht, und ob es als Hilfsmittel für die automatische Klassifikation genutzt werden kann.
Die Wikipedia ist eine Enzyklopädie, also eine strukturierte, möglichst umfassende Darstellung menschlichen Wissens in einer für den Alltagsgebrauch hinreichenden Ausführlichkeit. Wikipedia wird nicht von einer festen, bezahlten Redaktion, sondern von freiwilligen Autoren verfasst und aufgebaut.
Wikipedia ist eine kollaborative Plattform mit gewaltigem Wissenszuwachs. Diese damit verbundene Arbeit, kann alleine von keinem Unternehmen, Verband oder ähnlichem geleistet werden.
Eine redaktionelle Kontrolle der Inhalte wird in gewissen Grenzen vorgenommen, ergibt sich aber auch aus dem Prinzip der Zusammenarbeit, da sich in der Regel mehrere Fachleute eines Gebietes mit der Erstellung eines Artikels beschäftigen. Für Konfliktfäl-

le gibt es ein festgelegtes Reglement, das als letzte Konsequenz, die Sperrung bzw. das Entfernen eines Artikels vorsieht.

Für einzelne Wissensgebiete existieren Portale. Diesen Portalen sind Themen zugeordnet, die im Wikipedia als Kategorien bezeichnet werden. Das Anlegen von Kategorien unterliegt nur einigen wenigen Bedingungen und dient im Wesentlichen dazu bestimmte Bereiche oder Themen zu einer Begrifflichkeit zusammenzufassen. Die Strukturierung der Themen reicht durchschnittlich bis zu einer Tiefe von 4 Ebenen. Artikel können beliebigen Kategorien oder auch Portalen zugeordnet werden.

Ein Nachteil in der derzeitigen Version ist jedoch, dass die Hierarchiestruktur aus der Artikelsicht nur schwer erkennbar ist, da nicht der gesamte Hierarchienbaum sichtbar ist, sondern nur lediglich die zugeordnete Kategorie. Dadurch ist nicht deutlich sichtbar wo genau die Einordnung in der Struktur vorgenommen wurde. Das spiegelt sich auch in der zugrunde liegenden Datenbankstruktur wieder, bzw. die derzeitige Struktur lässt keine bessere Darstellung zu.

Darstellung der für die Hierarchie relevanten Tabellen



Bsp: Mit einem einfachen SQL Befehl können alle Artikel, die der Kategorie Informatik zugeordnet sind, gefunden werden.

```
SELECT cur_title  FROM cur, categorylinks WHERE cur_id = cl_from and
cur_namespace="de" and cl_to like 'Informatik'
```

Es werden jedoch nicht die Artikel gefunden, die einer oder mehreren Unterkategorien von ‚Informatik' zugeordnet sind. Das ist in der Regel aber der Fall, somit müsste für alle Kategorien im Portal Informatik die Abfrage durchgeführt werden, da die Beziehungen der Kategorien zueinander im zugrunde liegenden Datenbankschemata nicht berücksichtigt sind.

Ein Vorteil für eine Weiterverarbeitung der Daten in einer eigenen Applikation, ist die Möglichkeit Wikipedia komplett in XML Notation herunter zu laden. In regelmäßigen Abständen wird ein Abzug aller Artikel erzeugt und ebenfalls als Download unter http://download.wikimedia.org/wikipedia/de zur Verfügung gestellt. Dies ist ein großer Vorteil um darauf Anfragen lokal durchzuführen, bzw. die Daten auf beliebige Art und Weise zu verarbeiten.

Es gibt weiterhin die Möglichkeit SQL-Anfragen auf Wikipedia auszuführen, allerdings geht dies nur über das Browserinterface und die Ergebnisse werden via Mail gesandt.

Wikipedia liegt multilingual vor, allerdings ist die Entwicklung in anderen Ländern meist nicht so weit fortgeschritten, wie das in Deutschland der Fall ist. Auch gibt es kei-

ne Zuordnung, der Artikel untereinander. Ein Artikel, der in der deutschen Wikipedia vorhanden ist hat keinen Bezug zu einem eventuell vorhandenen anderssprachigen Artikel mit gleichem Inhalt. Die jeweiligen Wiki's sind voneinander vollkommen unabhängig.

Interessant für die Betrachtung ist der Kategoriebaum des Portals Informatik sein: [Wikipedia: WikiProjekt Informatik/KategorieBaum – Wikipedia]. Der Kategorienbaum zur Informatik besteht zum aktuellen Zeitpunkt aus 290 Kategorien, denen insgesamt 14095 Artikel zugeordnet sind (Stand 20. März 2006). Allerdings wiederholen sich einige Kategorien innerhalb des Kategorienbaums, und diesen Kategorien innerhalb unterschiedlicher Hierarchiestufen sind allerdings die gleichen Artikel zugeordnet. Das heißt, dass gleiche Begriffe in verschiedene Kategorien vorkommen, aber keine unterschiedliche Bedeutung hinsichtlich ihres Umfelds haben. Dies widerspricht eigentlich der Logik eines hierarchischen Systems, ist aber letztendlich auf die Entstehungsweise von Wikipedia zurückzuführen, da die Kategorien sozusagen ‚historisch gewachsen' sind.

Betrachtungen ob und wie Wikipedia als Datenquelle genutzt werden könnte, sind im Abschnitt Datenanreicherung beschrieben.

Gegenüberstellung von SWD, Wikipedia und CR-Classification

In der folgenden Tabelle soll versucht werden, die einzelnen Anwendungen bezüglich der Thesaurus Definition gegenüberzustellen.

Merkmalsvergleich konkreter Systeme			
Thesaurus Definition	**SWD**	**de.wikipedia.org**	**CR-Classification**
UF - used for	Schlagwort	Artikel-Titel	Node (CR_Deskriptor, CR_Label)
USE/SYN use synonym	S	Text	is composed by

BT - broader term	BT	Kategorien (Mächtigkeit N)	Parent Node
NT - narrower term	UB	Hyperlinks (vage)	Children Nodes
RT - related term		Hyperlinks (vage) (Siehe auch Verweise)	Is related to
TT - top term		Portal	Root Node

Für alle Systeme sind der Vorzugsdeskriptor, in Form von Schlagwort, Titel oder Klassenbezeichnung, sowie Vater-Kind-Beziehungen vorhanden. Die Nutzung von Synonymen wird in der CR Classification nur bedingt unterstützt, auch sind es nicht Synonyme im herkömmlichen Sinn, sondern eher näher beschreibende Begriffe, die spezifischer als die Klasse selbst sind und mögliche Ausprägungen der Klasse darstellen.
Für die Erfüllung der Thesaurus-Definition ist neben den Vorzugsdeskriptoren auch die Menge der Nicht - Vorzugs - Deskriptoren entscheidend. Diese Menge wird bei der CR Classification nicht erreicht. Für die Wikipedia kann man sagen, dass diese Bedingung erfüllt ist, allerdings ist hier die Strukturierung nicht sehr stringent und vor allem sind die Zuordnungen zu anderen Begriffen also Artikeln eher vage. In der Wikipedia ist die Kategorisierung zwar nicht sehr streng, dafür ist sie aber polyhierarchisch, was eine schöne Basis für eine Suche über verschiedene Aspekte ermöglicht. In der SWD ist die Hierarchie sehr flach, dafür aber die Menge der Schlagworte sehr hoch. Es ist schwer überhaupt eine Einteilung in Vorzugs- und Nicht- Vorzugs- Deskriptoren vorzunehmen. Ein Schlagwort hat maximal einen Oberbegriff eine beliebige Menge an Verweisen und Synonyme. Definitionen sind für Oberbegriffe vorhanden. Durch die schwache Strukturierung eignet sie sich nicht als Ordnungssystem. Die netzartige Struktur, die durch die gleiche Verwendung von Verweisen und Oberbegriffen entsteht ist nicht immer in sich geschlossen, was zu Inkonsistenzen bei einer Anwendung als Ordnungssystem führen würde.

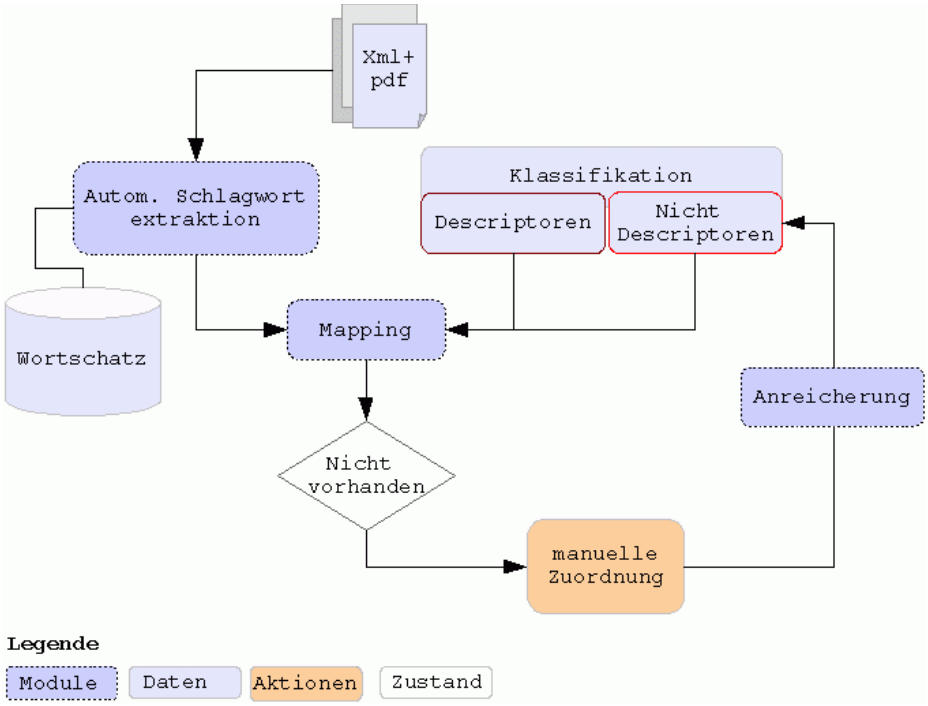
8 Entwurf eines Verfahrens

Der Entwurf des Verfahrens zur automatisierten Klassifikation gliedert sich in folgende wesentliche Punkte:

- Anwendung eines Basissystems, welches die Grundfunktionalitäten eines digitalen Repositories erfüllt.
- Erweiterung des Klassifikationssystems für die Anreicherung mit Deskriptoren.
- Ermittlung der relevanten Begriffe durch Terminologieextraktion
- Zuordnung der Begriffe und des zu Klassifizierenden Textes zu Kategorien der Fachklassifikation.

Einen groben Architekturüberblick zum Ablauf der Klassifikation eines Dokuments im Gesamtsystem veranschaulicht die folgende grafische Darstellung:

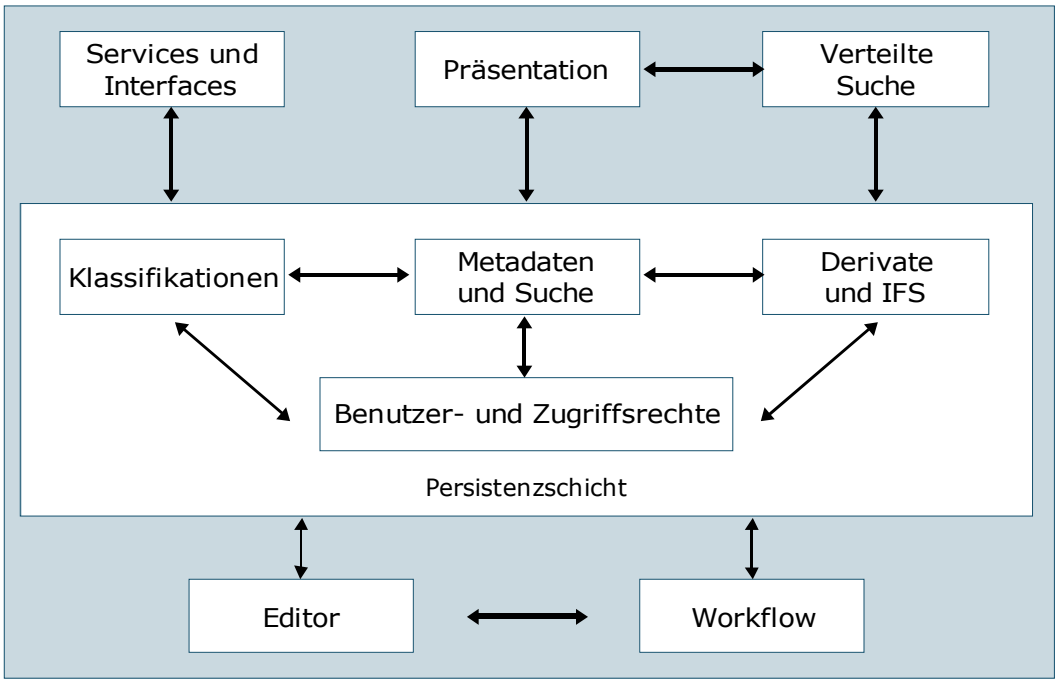
Entwurf Architekturmodell



Digitales Repository - MyCoRe

Als Basis soll das Open Source System MyCoRe genutzt werden. MyCoRe (My Content Repository) ist ein Kernsystem, das alle Grundfunktionen zur Nutzung digitaler Sammlungen enthält. Das MyCoRe Projekt gliedert sich zum einen in den Kern, welcher grundlegende Funktionalitäten und eine Programmierschnittstelle (API) bereitstellt und in die darauf aufsetzenden Anwendungen, welche neben einer eigenen Präsentation auch die Kern-Komponenten erweitern und/oder umgestalten können. Somit ist gewährleistet, dass MyCoRe in vielen Projekten als Datenbasis eingesetzt werden kann. Die wesentlichen Hauptkomponenten von MyCoRe sind die Speicherung der Daten in einem Digitalen Repository, die Metadaten- und Volltextsuche, die Klassifikationsmöglichkeiten, sowie eine Benutzerverwaltung, Editoren und Workflowmodule.

MyCoRe Architekturmodell



Basis für die Dokumentenspeicherung und Präsentation ist ein MyCoRe Datenmodell, das eine flexible Modellierung von Metadaten bei möglichst geringem Arbeitsaufwand für die Einzelanwendung gestatten soll. Es sind folgende Ansätze implementiert:
Ein MyCoRe-Metadaten-Objekt besteht aus den drei Teilen:

- **structure -** hier werden Informationen zur Einordnung des Datensatzes in eine logische Struktur festgehalten. Dies umfasst Verweise zu Eltern- und Kind-Metadaten-Sätzen. Weiterhin werden in diesem Abschnitt Referenzen zu den Metadaten (Derivaten) der eigentlichen digitalen Objekte gespeichert.
- **metadata -** in diesem Bereich werden die eigentlichen Metadaten gemäß dem Datenmodell abgelegt. Jeder MyCoRe-Metadaten-Typ kann hier als eine Liste von Elementen angegeben werden. Zu jedem Metadaten-Typ können beliebige Attribute definiert werden, Parameter zur Suchbarkeit des Datentyps und dessen Vererbung, sowie sprachspezifische Eigenschaften festgelegt werden.
- **service -** der letzte Abschnitt enthält Informationen zur Verwaltung des Metadaten-Satzes wie Erstellungsdatum u.a.

Ein Dokument besteht aus dem MyCoRe-Metadaten-Objekt und einer beliebigen Anzahl von Derivaten, zur Beherbergung der eigentlichen digitalen Objekte, wie z.B. den Volltext einer Dissertation. Es gibt einen Grundstamm an XML-basierten Datentypen, die zur Beschreibung beliebiger Daten genutzt werden können.

Es können mit Hilfe der Grundtypen, beliebige eigene Datenmodelle aufgebaut werden. Dokumente können beliebig Klassifikationen zugeordnet werden. Mithilfe dieser Basismodule und deren Funktionalität kann eine eigene MyCoRe Applikation aufgebaut werden, die dann um eigene Inhalte und Funktionen erweitert werden kann. [MYCO06]

Erweiterung des Klassifikationssysteme in MyCoRe

Bezüglich der Klassifizierung von Fachtexten ist das MyCoRe - Klassifikationssystem wesentlich und soll aus diesem Grunde hier näher erläutert werden.

Um eine bessere Strukturierung von Informationen, besonders in großen Datenbeständen zu erreichen, wurden in allen Bereichen der Informationsverarbeitung Klassifikationen eingeführt. Dabei kann es sich um sehr allgemeine, wie der Herkunft oder eines Dateiformat-Typs, oder aber fachspezifische Klassifikationen für eng begrenzte Wissenschafts- oder Sachgebiete, handeln.

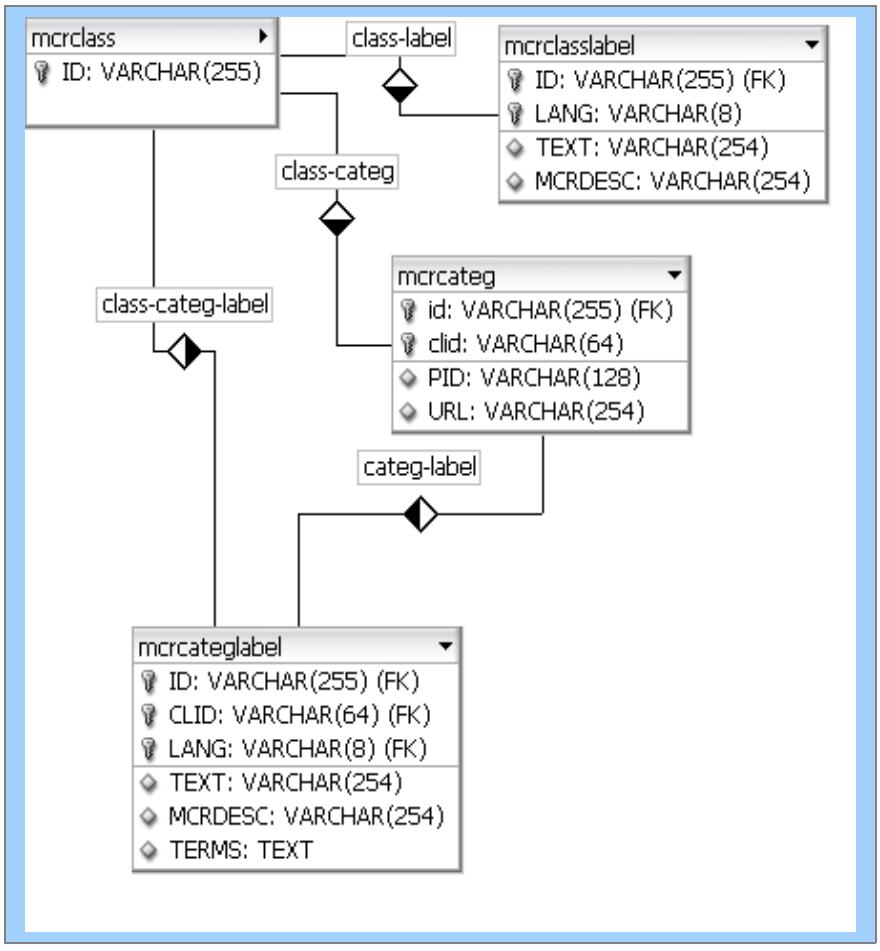
MyCoRe implementiert ein Datenmodell mit folgenden Eigenschaften:

- Abbildung von hierarchischen Strukturen einer Klassifikation.
- Eindeutige Identifizierung einer Kategorie über mehrstufige Identifier.
- Berücksichtigung der Mehrsprachigkeit für Kategoriebezeichnungen
- Speicherung von Text-Beschreibungen zu den einzelnen Kategorien.
- Speicherung einer URL zu jeder einzelnen Kategorie.

Jede Klassifikation ist in XML Notation beschrieben und kann mit den für XML vorhandenen Werkzeugen bearbeitet werden. Eine Überprüfung der Syntax und Semantic er-

folgt über XML-Schema. Für die Einbettung des Verfahrens zur automatischen Klassifikations in MyCoRe habe ich das Klassifikationsmodell erweitert, um die Terme zur Beschreibung der Kategorien aufnehmen zu können.

Modell: Klassifikationen



Für jede Klassifikation existiert ein Eintrag in der Tabelle `mcrclass` sowie für jede sprachliche Ausprägung ein Eintrag in der Tabelle `mcrclasslabel`. Jede Kategorie in der Tabelle `mcrcateg` besitzt eine eindeutige ID und ist einer Klassifikation zugeordnet. Sprachabhängig können Kategorien beliebige Label besitzen. Jeder Eintrag in der Tabelle `categlabel` enthält die Bezeichnung der Kategorie, eine Beschreibung und das Attribut `TERMS`, in dem alle Terme, die der Kategorie zugeordnet sind separiert abgelegt werden können. Eine Klassifikation kann beliebig viele Kategorien besitzen. Kategorien können beliebig viele Unterkategorien besitzen.
Die CR Classification, die ich über eine XSL-Transformation in das MyCoRe Klassifikationsmodell konvertiert habe, wird durch folgende XML Notation dargestellt. (Auszug)

```
<mycoreclass  ID="mycore_class_00000100"
xsi:noNamespaceSchemaLocation="MCRClassification.xsd">
 <label description="The ACM Computing Classification System     [1998
Version] Copyright 2006, by the Association for      Computing…" />
 <categories>
  <category ID="I.7">
    <label xml:lang="en"
            text="I.7 DOCUMENT AND TEXT PROCESSING" description="" />
      …
      <category ID="I.7.2">
       <label xml:lang="en" text="I.7.2 Document Preparation"
           description="H.4|H.5">
           Desktop publishing | Format and notation |
           Hypertext/hypermedia |Index generation | Languages and
           systems |Markup languages | Multi/mixed media |
           Photocomposition/typesetting | Scripting languages |
           Standards |
       </label>
      </category>
           …
  </category>
          …
 </categories>
</mycoreclass>
```

Die Terme, die beim Verfahren des Klassifizierens, durch die Termextraktion gefunden und der Kategorie zugeordnet werden, werden im Content des zugehörigen Labels abgelegt. Im Attribut *description* werden, soweit vorhanden, die relevanten Kategorien, die durch [is related to] in der CR Classification beschrieben werden, abgelegt.

Die ursprüngliche CR Classification liefert in ihrer XML Notation schon einige wenige Terme, die im <composed by> Block der XML Notation der CR Classification abgelegt sind. Diese Terme wurden bei der Transformation übernommen.
Um eine automatische Klassifikation mit der CR Classification realisieren zu können, ist eine Anreicherung der Klassen mit Synonymen und weiteren Deskriptoren notwendig. Die Anreicherung kann einerseits durch einen Lernprozess sukzessiv vorgenommen werden, andererseits könnte man versuchen mit geeigneten Basisdokumenten zunächst eine passable Startmenge an Metainformation aufzubauen. In meinem Entwurf und dessen Umsetzung habe ich die Variante des Lernens bevorzugt, da die Ermittlung einer geeigneten, genügend großen Begriffsquelle vor allem auch aus Zeitgründen nicht möglich war.

Terminologieextraktion und Klassifikation

Die Klassifikation der Dokumente soll durch zwei wesentliche Komponenten erreicht werden. In einem ersten Schritt werden aus den zu klassifizierenden Dokumenten die Fachtermini extrahiert. Die Ermittlung der für den Text relevanten Schlagworte soll mit Hilfe des *Indexers* erfolgen, ein Tool das auf statistischen und linguistischen Verfahren, wie in Kapitel 3 Textindizierung und Merkmalsextraktion vorgestellt, basiert. Der *Indexer,* den ich dankenswerterweise zur Nutzung in der vorliegenden Arbeit zur Verfügung gestellt bekam, wurde in der Abteilung Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig im Rahmen verschiedener Forschungsprojekte entwickelt und wird ständig weiter bearbeitet. Dabei handelt es sich um ein Tool, das durch die Kombination verschiedener Verfahren, wie statistische Differenzanalyse oder musterbasierte reguläre Ausdrücke über Part Of Speech Tags (POS), also das Kategorisieren von Einheiten (Wörtern) in Verbünden (Sätzen) mithilfe linguistischen Wissens, versucht, Fachterminologie aus Fachtexten zu extrahieren.

Das zu klassifizierende Dokument ist der zu analysierende Textkorpus, der wesentlich größere Referenzkorpus wird aus Daten des Projektes Deutscher Wortschatz (http://wortschatz.informatik.uni-leipzig.de/) gebildet. Auf dem Analysekorpus wird zunächst eine Differenzanalyse bezüglich des Vorkommens von Begriffen und deren Häufigkeit im Referenzkorpus ausgeführt. Dabei werden die im Analysetext am häufigsten auftretenden Begriffe und Terminologien ermittelt, die inhaltlich relevant sein könnten. Für die Ermittlung kann sich auf Nominative, Adjektive oder Verben beschränkt werden. Zur Ermittlung von Schlagworten aus Fachtexten, habe ich die Auswahl auf Nominative beschränkt. Neben der Frequenz der Begriffe ist vor allem das Maß für die Relevanz von Bedeutung.

Das Signifikanzmaß kann, für gemeinsprachliche Texte z. B. über den likelihood Ratio Test (LR) ermittelt werden. Der LR ist ein relativ allgemein einsetzbares Verfahren zum Vergleich von Modellen auf der Grundlage des Maximum-Likelihood Verfahrens. Verglichen werden jeweils zwei Modelle: Ein Ausgangsmodell, welches ein oder mehrere Modellparameter enthält und ein Vergleichsmodell, in welchem einem oder mehreren dieser Parameter Restriktionen auferlegt wurden. Im Fall der Analyse des Fachtextes bezüglich des Vergleichskorpus, also der Vergleich des Auftretens eines Wortes aus dem Fachtext im gemeinsprachlichen Korpus. Ausgegangen wird dabei von der so genannten Nullhypothese. Diese Prüfung besagt, ob im zu analysierenden Modell das zu prüfenden Element signifikant öfter vorkommt, als rein durch Zufallsschwankungen zu erwarten wäre. Beim LR ist jede Prüfung unabhängig von den vorangegangenen Prüfungen. Für Worte die in einem Kontext zueinander stehen, wäre diese statistische Nähe zwar interessant, wird aber nicht berücksichtigt. Weiterhin ist die Wahrscheinlichkeit ein positives Testergebnis zu erhalten für jede Prüfung gleich, wobei auch diese Annahme nicht ganz stimmt, für die Gesamtheit aller Test der Worte sich jedoch ausgleicht. Insgesamt gilt für Worte und ihr wahrscheinliches Auftreten in Texten die Binominalverteilung.

Eine andere Möglichkeit besteht darin, den Häufigkeitsquotienten eines Wortes, also den Quotienten aus relativer Häufigkeit eines Wortes im Fachtext und seiner relativen Frequenz im Vergleichskorpus zu ermitteln [WITS051]. Alle Häufigkeitswerte weisen

den Worten einen numerischen Wert zu, nach dem diese geordnet werden können. Unterhalb eines Schwellenwertes wird die gewonnene Liste abgeschnitten. Die Verfahren kann man sowohl auf Indexterme, als auch auf Buchstaben-N-Gramme anwenden. Bei der Verwendung von N-Grammen, wird der Schwellenwert durch eine Mindestanzahl von Wörtern in denen ein N-Gramm auftritt festgelegt.
Nach [WITS051] ist es je nach Art des Textes davon abhängig, welches Prüfverfahren bessere Ergebnisse liefert. Nach seinen Tests liefert die Nutzung des Häufigkeitsquotienten für Fachtexte bessere Ergebnisse als die Nutzung des likelihood Ratio Prüfverfahrens. Für gemeinsprachliche Texte ist dies nicht so, da der Häufigkeitsquotient seltene Termen stärker bewertet.
Für mein Verfahren zur automatischen Klassifikation, soll die Termextraktion des *Indexers* mit der Signifikanzbestimmung über den Häufigkeitsquotienten genutzt werden. Als Schwellenwert habe ich einen sehr niedrigen Wert angegeben, da die Frequenz sehr von der Größe des zu analysierenden Textes abhängt. Ist der Schwellenwert zu hoch gewählt, werden bei einem sehr kurzen Text eventuell zu wenige Terme extrahiert. In der prototypischen Implementation habe ich als Schwellenwert die Anzahl zwei festgelegt, schneide aber die Ergebnismenge jeweils für gefundene Ein- und Mehrwortterme bei der Anzahl zehn ab.
Als Ergebnis bekommt der Anwender eine Liste von jeweils zehn Einwort- und Mehrworttermen vorgeschlagen, die er zur Klassifizierung des Dokumentes übernehmen kann.

Zunächst habe ich den *Indexer* in die eigene MyCoRe Applikation integriert, um die Ergebnisse der Textextraktion dem Anwender zur weiteren Verarbeitung zur Verfügung zu stellen.
Um die Umsetzung etwas besser zu veranschaulichen, soll der Gesamtprozess zur Klassifikation eines Fachtextes an einem Beispiel veranschaulicht werden. Dabei habe ich einen Fachtext aus der Dokumentenmenge, die ich automatisiert klassifiziert habe, ausgewählt, der bereits manuell mit Schlagworten und Klassifikation versehen wurde:

Titel des Fachtextes: *Datenverteilung in Peer-to-Peer Overlay-Netzwerken*

Zusammenfassung: *In dieser Diplomarbeit wurde ein mobiles Informationssystem (MIS) auf Basis eines dezentralisierten, unstrukturierten Peer-to-Peer Overlay Netzwerkes entwickelt. Es werden Strategien zur Verteilung von Daten im Netzwerk (Replikation) und Probleme bei Änderungsoperationen (Updates) diskutiert. Geeignete Ansätze zur Umsetzung der Datenverteilung und der Updates wurden gewählt und auf das System angepasst. Um die eingeführten Konzepte zu validieren, wurde ein Prototyp entwickelt.*

Manuell vergebene Schlagworte: *Dezentralisiert, unstrukturiert, Peer-to-Peer Overlay Netzwerk, Ad-Hoc, Replikation, Distributed Hash Tables, Updates*

Manuell vergebene CR Classification: *C.2.1 Network communications, C.2.4 Distributed applications, E.2 Hash-table representations, H.2.4 Distributed systems, H.3.2 Information Storage, H.3.4 Information networks*

[POHL2004]

Die folgende Abbildung zeigt die Ergebnisse der Terminologieextraktion:

Beispiel: Termextraktion in MyCoRe

Ergebnis | automatische Schlagwortextraktion

TestDiplomarbeit

C:\atLibriJSP\export\autoclassification\da-pohl.txt

Einwortterm	Häufigkeit	Signifikanz	Auswahl
Knoten	301	3923.	☐
Objekt	291	2707.	☐
Netzwerk	201	2304.	☑
Replikat	119	2250.	☐
Replikation	68	1221.	☑
Tree	75	1200.	☐
Nachricht	139	1122.	☐
Peer	71	954.2	☐
Tapestry	43	853.1	☐
Algorithmus	52	778.3	☐

Mehrwortterm	Häufigkeit	Signifikanz	Auswahl
D Tree	69	69.0	☐
mobil Informationssystem	49	49.0	☐
Peer to	37	37.0	☐
to Peer	31	31.0	☐
Peer to Peer	31	31.0	☑
Peer Netzwerk	22	22.0	☐
to Peer Netzwerk	21	21.0	☐
Top K	15	15.0	☐
Ping Nachricht	15	15.0	☑
Overlay Netzwerk	13	13.0	☑

Arbeit abbrechen | Für Klassifikation übernehmen

Zum jeweiligen Term werden die Frequenz, also die Anzahl des Auftretens des Terms im Gesamtdokument und seine Signifikanz dargestellt. Zum Beispiel der Begriff ‚Replikation' tritt im analysiertem Dokument mit einer Signifikanz von 1221 auf, d.h. die Wahrscheinlichkeit des Auftretens im Analysetext ist bezüglich der Wahrscheinlichkeit des Auftretens im Vergleichskorpus um diesen Wert höher.

In zweiten Schritt müssen die gefundenen freien Terme einer oder mehreren Kategorien der CR Classification zugeordnet werden. Durch einen einfachen automatischen Vergleich von Termen und Deskriptoren der Klassifikation ist dies nicht möglich, zumindest nicht, bevor die Klassifikation nicht mit einer hinreichend großen Menge an Termen angereichert wurde, die dann als Deskriptoren der jeweiligen Kategorie dienen.
Dies hat folgende Gründe:

- der Vorzugsdeskriptor selbst, dient meist der Beschreibung eines Themas
- der Vorzugsdeskriptor ist nur in seinem Kontext, innerhalb der Hierarchie sinnvoll.
- Vorzugsdeskriptoren dienen auch der Abgrenzung zu anderen Deskriptoren, als eine Art Antideskriptor, der zum Klassifizieren all derjenigen Texte dient, die zwar in einer ausgewählten Hierarchie anzusiedeln sind, aber keinem der vorhandenen Deskriptoren zugeordnet werden können. Typischerweise sind das in der CR Classification die [[N].0 Generell] Deskriptoren.
- der gefundene Term kann in verschiedenen Kontexten verwendet werden.
- der Term kann überbewertet sein.

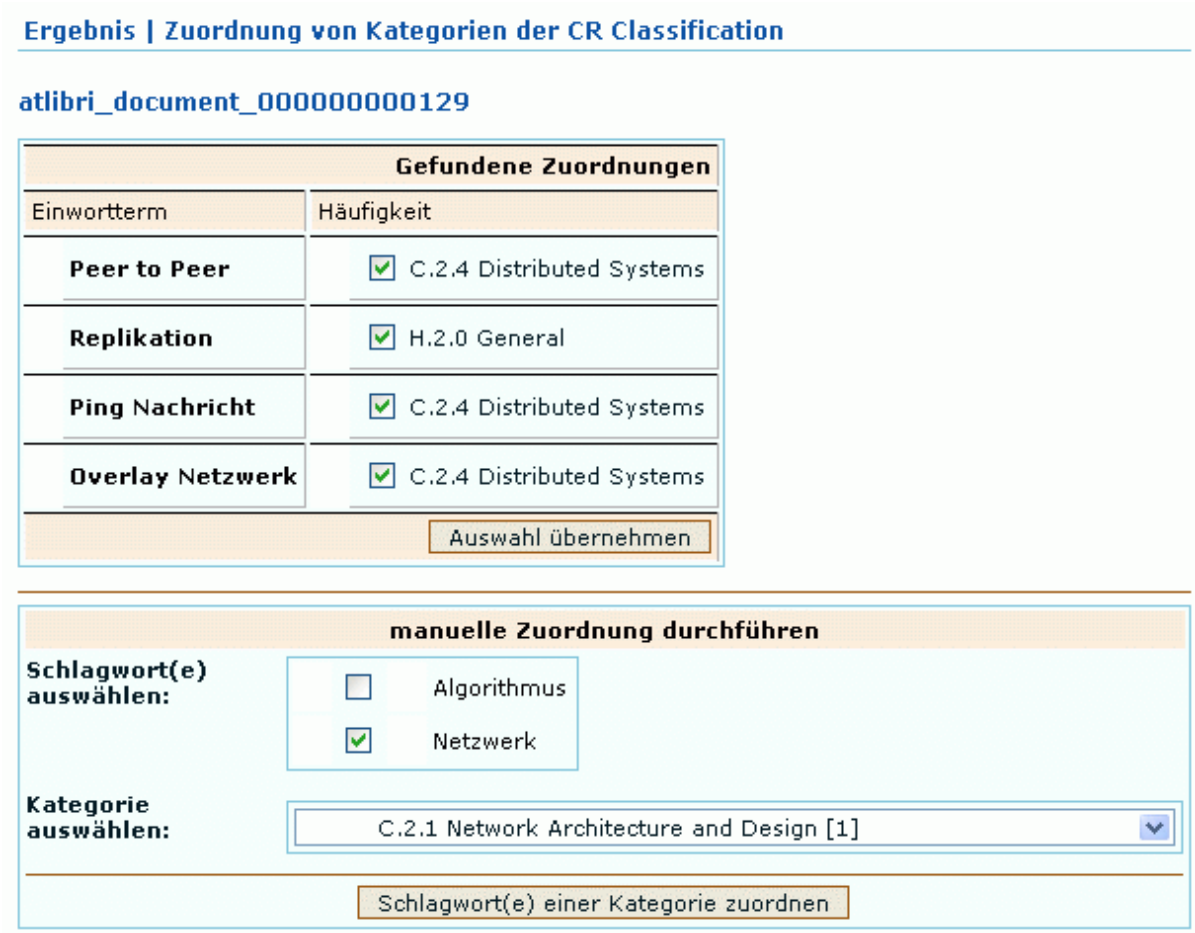
Um die unerwünschten Terme nicht automatisch mit zu übernehmen, wird eine Auswahl angeboten, die es ermöglicht, nur diejenigen Terme zu wählen, die aus Nutzersicht sinnvoll erscheinen.
Für jeden ausgewählten Term aus dem Extraktionsergebnis wird eine Suche über alle Kategorien der Klassifikation durchgeführt, ob Kategorien existieren, in denen der TERM bereits als Deskriptor vorkommt.

Dabei gibt es folgende Ergebnisvarianten pro gefundenen Term:

- Es existiert für einen Term kein Treffer: Der Term ist in der gesamten Klassifikation noch nicht vorgekommen der Anwender bekommt eine Auswahl aller Kategorien zur Klassifikation des Terms.
- Es existiert genau 1 Treffer, also der Term wurde in genau einer Kategorie gefunden. Dieser Fall ist der einfachste. Es erfolgt eine Anzeige von Term und Kategorie, deren Zuordnung wahlweise übernommen werden kann.
- Der Term kann bereits mehreren Kategorien zugeordnet worden sein. Mehrere Terme können einer Kategorie zugeordnet werden. Das Ergebnis wird als Vorschlag dem Anwender zur Verfügung gestellt. Ein TERM kann in verschiedenen Kategorien auftauchen, da er im Zusammenhang mit anderen Termen einen Inhalt beschreibt, der durch genau die Zusammenstellung dieser Kategorien dargestellt wird. Sind die Treffer in verschiedenen Kategorien so muss die Beziehung der Kategorien bezüglich der Hierarchie berücksichtigt werden. Treffer in verschiedenen Hierarchiezweigen müssen nebeneinander dargestellt werden. Treffer innerhalb eines Hierarchiezweiges könnte man auf den jeweils obersten Hierarchieknoten beschränken, da man damit eine allgemeinere Aussage treffen kann.

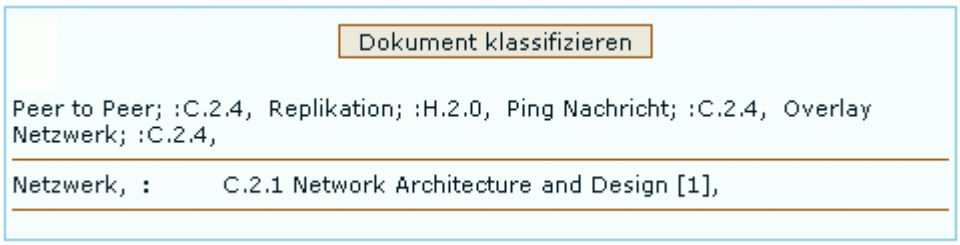
In der Anwendung wurden die möglichen Ergebnisvarianten so dargestellt, dass immer eine Abwahl oder Auswahl der Zuordnung TERM – Kategorie realisierbar ist.

Beispiel: Klassifikation Termen zuordnen



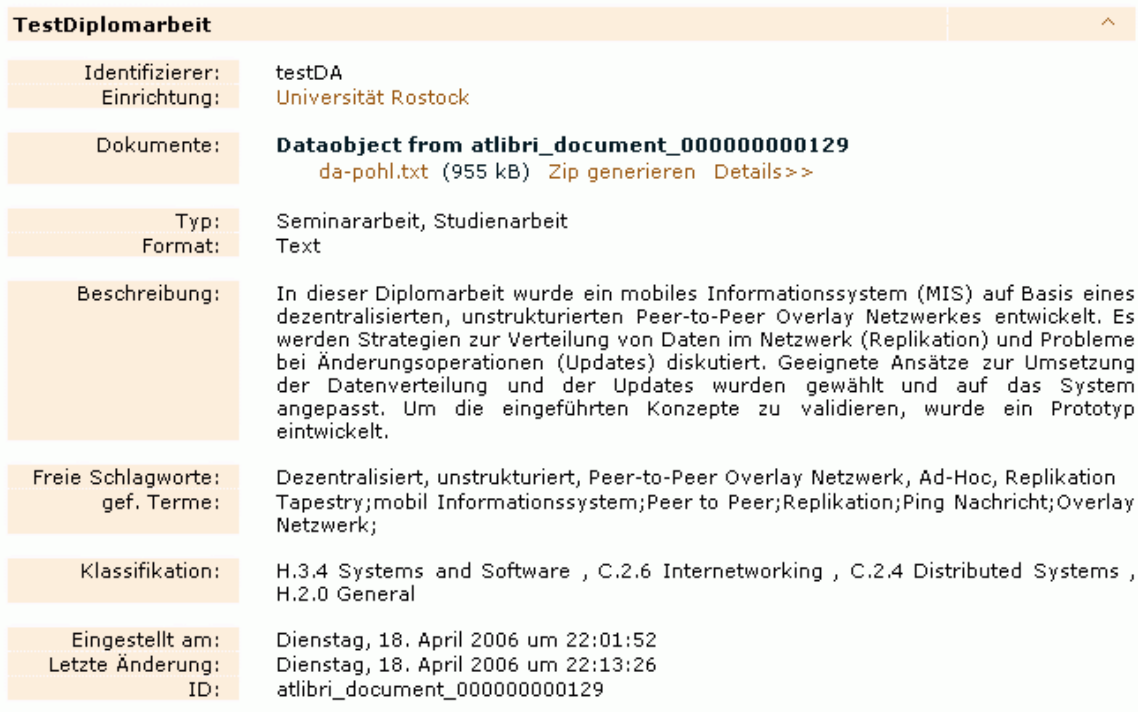
In der oberen Tabelle, sind die Terme aufgelistet, die der Nutzer ausgewählt hat und die schon innerhalb einer bzw. mehrerer Kategorien gefunden wurden. Im unteren Teil, sind die Terme aufgelistet, die noch manuell zugeordnet werden müssen. Für jeden Term kann die Übernahme und Zuordnung zur ausgewählten Kategorien bestätigt werden. Im Ergebnis sind dann die Zuordnungen im unteren dritten Teil des Interaktionsdialogs zusammengefasst.

Beispiel: Dokument klassifizieren



Damit werden die Klassifikation des Dokuments und die Zuordnung von Termen zur Klassifikation für das analysierte Dokument abgeschlossen. Das Ergebnis ist in der Detailansicht zur Darstellung der Metadaten sichtbar.

Beispiel Metadaten des Dokuments



TestDiplomarbeit

Identifizierer:	testDA
Einrichtung:	Universität Rostock
Dokumente:	**Dataobject from atlibri_document_000000000129** da-pohl.txt (955 kB) Zip generieren Details>>
Typ:	Seminararbeit, Studienarbeit
Format:	Text
Beschreibung:	In dieser Diplomarbeit wurde ein mobiles Informationssystem (MIS) auf Basis eines dezentralisierten, unstrukturierten Peer-to-Peer Overlay Netzwerkes entwickelt. Es werden Strategien zur Verteilung von Daten im Netzwerk (Replikation) und Probleme bei Änderungsoperationen (Updates) diskutiert. Geeignete Ansätze zur Umsetzung der Datenverteilung und der Updates wurden gewählt und auf das System angepasst. Um die eingeführten Konzepte zu validieren, wurde ein Prototyp eintwickelt.
Freie Schlagworte:	Dezentralisiert, unstrukturiert, Peer-to-Peer Overlay Netzwerk, Ad-Hoc, Replikation
gef. Terme:	Tapestry;mobil Informationssystem;Peer to Peer;Replikation;Ping Nachricht;Overlay Netzwerk;
Klassifikation:	H.3.4 Systems and Software , C.2.6 Internetworking , C.2.4 Distributed Systems , H.2.0 General
Eingestellt am:	Dienstag, 18. April 2006 um 22:01:52
Letzte Änderung:	Dienstag, 18. April 2006 um 22:13:26
ID:	atlibri_document_000000000129

9 Auswertung der Testreihe und Vergleich der manuellen und der semiautomatischen Klassifikation

Für die Testreihe wurden insgesamt ca. 30 im Internet frei verfügbare deutschsprachige Dissertationen der TU München klassifiziert, zu denen es schon eine manuelle Zuordnung von Schlagworten und Fachklassifikationen gab. Die Dokumente standen über die Sammlung Digitale Dissertationen (http://mediatum.ub.tum.de) zur Verfügung. Für

eine umfassende Gegenüberstellung sind neben den Ergebnissen der Klassifikationsverfahren auch mindestens die Metadaten wie Titel, Untertitel, Abstrakt und Schlagworte wesentlich. Aufgrund des Umfangs möchte ich hier aber nur kurz eine Gegenüberstellung der Ergebnisse mit den Daten zu Titel, Klassifikation und Schlagworten einiger analysierter Dokumente aufzeigen.

Titel 1	**Objektorientierte Daten- und Zeitmodelle für die Echtzeit-Bildfolgenauswertung**
SWD	Bildfolge , Modellierung , Objektorientierung , Echtzeitsystem; Bildfolge, Auswertung , Echtzeitsystem; Bildfolgenverarbeitung , Echtzeitverarbeitung , Objektorientierung; Fußball , Autonomer Roboter
Systematik	DAT 760d - Bildverarbeitung; Computer Vision; Maschinelles Sehen DAT 260d - Echtzeitverarbeitung
aus Termextraktion übernommen	Sequenz; Datensequenz; physikalisch Sensor; RoboCup; Sequenz; Sensor; logisch Sensor;
CRC	I.2.9 Robotics; I.4.8 Scene Analysis

Titel 2	**Systematische Analyse und Konstruktion integrierter Sicherheitsarchitekturen für mobile verteilte Systeme**
SWD	Betriebssystem , Verteiltes System , Computersicherheit , Systemmodell
Systematik	DAT 423d - Zuverlässigkeit von Betriebssystemen; DAT 461d - Datensicherheit in Rechnernetzen
aus Termextraktion übernommen	Sicherheitsanalyse; Schwachstelle; Integriert Sicherheitsanalyse; System; mobil System; IT System; Betriebssystem; IT Sicherheit; Sicherheitsanalyse; IT Sicherheit; Integriert Sicherheitsanalyse;
CRC	D.4.0 General , D.4.6 Security and Protection , K.6.5 Security and Protection

Titel 3	**Konzeption und Anwendung einer Customer Service Management Architektur**
SWD	Management , Schnittstelle , Architektur (Informatik); Organisationsmodell , Management; Informationsmodell , Management; Kommunikationsmodell , Management
Systematik	DAT 250d - Rechnernetze; Verteilte Datenverarbeitung; DAT 899d - Sonstiges; DAT 060d - DV-Management; Rechenzentrum; WIR 546d - Management-Informationssystem; WIR 570d - Information (Informationsfluss, -arten, -wege)
aus Termextraktion übernommen	CSM; Operativ CSM; Management; Service Management; CSM Schnittstelle; Schnittstelle; CSM Architektur; Service Management; CSM Architektur
CRC	H.4.1 Office Automation K.6.1 Project and People Management ,

Titel 4	**Ein Konzept zur Lastverwaltung in verteilten objektorientierten Systemen**

Freie Schlagworte	verteilte Systeme, objektorientiert, Lastverwaltung, CORBA
Systematik	n. v.
aus Termextraktion übernommen	Lastbewertung; Lastverteilung; dynamisch Bindung; verteilt System; Lastverwaltung; CORBA;
CRC	C.2.4 Distributed Systems, H.4.3 Communications Applications, H.5.3 Group and Organization Interfaces

Titel 5	**Ansichtsbasierte Objekterkennung mit Hilfe optimierter Musterbäume**
SWD	Serviceroboter , Greifer , Videobild , Objekterkennung; Videobild , Bildfolge, Objekterkennung , Objektmodell , Entscheidungsbaum, Klassifikator (Informatik); Gesicht, Videobild , Bildfolge, Objekterkennung
Systematik	DAT 815d - Robotik; DAT 760d - Bildverarbeitung; Computer Vision; Maschinelles Sehen
aus Termextraktion übernommen	Objekterkennung; Videosequenz; Entscheidungsbaum
CRC	I.2.0 General, I.2.10 Vision and Scene Understanding

Titel 6	**CSCW in der Bioinformatik: Ein objektorientiertes Groupwaresystem zur Unterstützung in der Gen- und Genomanalyse**
SWD	Computer supported cooperative work , Genanalyse
Systematik	DAT 612d - CSCW; Dialogsysteme; Transaktionssysteme; BIO 110d - Biometrie, Biomathematik, Datenverarbeitung in der Biologie
aus Termextraktion übernommen	CSCW System; Groupwaresystem; CORBA; CSCW; persistent Objekt; Informationsraum; Awareness Information; gemeinsam Informationsraum; persistent Speicher;
CRC	D.2.11 Software Architectures , H.4.3 Communications Applications, H.5.3 Group and Organization Interfaces

Titel 7	**Parallele Anfrageverarbeitung in multidimensionalen Array Datenbanksystemen**
SWD	Multidimensionales Datenbanksystem , Paralleles Datenbanksystem , Abfrageverarbeitung
Systematik	DAT 416d - Dialogbetrieb; DAT 516d - Parallelprogrammierung; Petrinetze; DAT 655d - Datenbanksysteme
aus Termextraktion übernommen	Array Daten; multidimensional Array; Parallel Verarbeitung; Anfrage; DBM
CRC	H.2 DATABASE MANAGEMENT, H.2.0 General, H.2.4 Systems

Zusammenfassend können einige allgemeine Aussagen zu den Ergebnissen getroffen werden. In keinem Fall gibt es eine vollkommene Deckungsgleichheit hinsichtlich der Klassifikation, unabhängig davon, ob unterschiedliche Klassifikationen verwendet wurden. Je mehr Abkürzungen verwendet wurden, desto besser sind die Klassifikationsergebnisse ausgefallen, die ja auf der Termextraktion fußen. Da Abkürzungen in Fachtexten für Fachtermini meist sehr relevant sind, kommt es hier zu recht guten Ergebnissen.

Vor allem Abkürzungen, wie CORBA, CSCW, XML, HTML oder DBM werden sofort gefunden und können zur Klassifizierung gut genutzt werden. Oft sind die extrahierten Terme spezifischer, als die dem Dokument manuell zugeordneten Schlagworte. Das ist in letzter Hinsicht für den Anwender als Fachmann günstig, für einen Laien weniger gut. Allerdings spiegelt sie ja auch das Spezifikationsniveau des Textes wider, insofern sind sie als freie Schlagworte für eine Recherche im Bestand des digitalen Repositories sehr gut geeignet.

Insgesamt liefert die Termextraktion für Fachtexte recht gute Ergebnisse, die sowohl für eine Recherche, als auch für die Anreicherung der Klassen als weitere Deskriptoren geeignet sind.

Bei der gesamten Tätigkeit des Klassifizierens, wurde ein Vorschlag zur Klassifikation vom System gegeben, der auch übernommen werden konnte (CORBA). Das ist zunächst recht dürftig, ersparte die Termextraktion aber zumindest die Arbeit, den Text zu analysieren und die wesentlichen Sinnträger herauszufinden. Da das ganze System auf der vorhandenen Menge an bereits klassifizierten Dokumenten und den damit verbundenen Fachtermen aufbaut, ist auch nur ab einer größeren Menge an klassierten Texten ein Automatisierungseffekt erkennbar. Eine Automatisierung könnte erkennbar werden, wenn jeder Klasse Terme aus ca. 10-20 Dokumenten zugeordnet sind. Da Fachtexte in der Regel mehreren Klassen einer Fachklassifikation zuzuordnen sind, benötigt man mindestens 30-40 Fachtexte eines Themenkomplexes. Es wäre sinnvoll in einer weiterführenden Arbeit zu untersuchen, ob eine Automatisierung dann auch wirklich erkennbar ist. Unabhängig von der Dokumentenmenge, die bereits klassifiziert wurde, wird es immer Terme geben, die neu hinzukommen und wieder einer Klasse zugeordnet werden müssen. Darin spiegelt sich die Weiterentwicklung des Fachgebietes wieder.

Ausblick, Probleme, weitere Variationen

Ein wesentliches Problem ist das vollständige automatisierte Klassifizieren bei Dokumenten mit übergreifenden Themen. Wie in der Auswertung zu sehen wurden Titel 3 und Titel 6 manuell Klassifikationen aus einem anderen Fachgebiet zugeordnet. Hier versagt eine einzelne Fachklassifikation natürlich, da wenn überhaupt nur über Klassen wie ‚Sonstiges…' oder ‚Anwendungen…' eine allgemeine Zuordnung vorgenommen werden kann. Dokumente mit übergreifenden Themen können mit einer Fachklassifikation nicht vollständig klassifiziert werden.
Ein weiterer Punkt, der bis hier gar nicht betrachte wurde, ist die Pflege der Klassifikation selbst. Wann sollten Deskriptoren einer Klasse entfernt werden, bzw. sollten sie überhaupt entfernt werden? Wie geht man mit einem Bedeutungswandel oder Zuordnungswandel von Termen innerhalb eines Kontext um?
Weiterhin könnte man darüber nachdenken, auch die Wichtung der Terme als Maß für die Relevanz der Deskriptoren zu berücksichtigen. Dieser Sachverhalt wird erst bei einer großen Menge an Deskriptoren innerhalb einer Klasse interessant, wenn vielleicht eine zu häufige Übereinstimmung mit Termen des analysierten Dokuments und Deskriptoren der Klassen auftritt.
Zur Verbesserung der Relevanz der Ergebnisse der Termextraktion ist die Ausnutzung der Metadaten eines Dokumentes sicher sinnvoll und innerhalb eines Digitalen Repositories, wie MyCoRe auch relativ einfach umsetzbar.

Nutzung der Metadaten von Online - Dokumenten

Oft liegen Elektronische Dokumente nicht nur in Form eines digitalen Objektes vor, sondern neben der eigentlichen digitalen Ressource, werden die beschreibenden Daten des Dokuments zusätzlich aufgeführt. Der Vorteil dabei ist, dass sich hier im Allgemei-

nen einer Auszeichnungssprache, wie z.B. XML bedient wird und Auszeichnungssprachen eine Syntax, eine Grammatik und eine Semantik besitzen.
Diese beschreibenden Daten - Metatags ermöglichen einen Zugriff auf semantisch strukturierte Informationen. Damit ist ein semantisch strukturierter Zugriff auf Teile von Dokumenten möglich und die Metatags können für die Interpretation der textuellen Informationen genutzt werden. Zum Beispiel kann man davon ausgehen, dass ein Begriff, soweit es sich um einen Nominativ handelt, der im Abstract oder im Titel eines Dokumentes auftaucht, eine größere Bedeutung hat, als ein Begriff aus dem Volltext.
Die im Entwurf des Verfahrens verwendeten Metadaten sind am Dublin Core Standard(DC) angelehnt [DCMI04]. Der DC Standard definiert eine Mindestmenge an Tags, die der Beschreibung eines Dokumentinhalts dienen.

Es sollten mindestens die folgenden Tags genutzt werden können:

Tag	**Beschreibung**
dc.title	Titel des Dokumentes
dc.subtitle	Untertitel
dc.description type={abstract|tableOfContents}	Beschreibung und /oder Inhaltsverzeichnis
dc.coverage	Themenumschreibung

Die Metadaten werden neben dem Volltext in der gleichen Weise indiziert, und auf Grund des geringeren Umfangs der textuellen Information in den Metadaten, wird auch die Relevanz der Terme gegenüber den Termen aus dem Volltext höher.

Nach Anwendung der Termextraktion erhält man eine Liste von Termen, sowohl für die Metadaten als auch für den Volltext. Beide 'Schlagwortlisten' werden für die weitere Verarbeitung zusammengefasst, die Relevanz der übereinstimmenden Terme könnten addiert werden und die relevantesten Terme dann als Deskriptoren zur Anreicherung der Klassen der Klassifikation dienen.

Datenanreicherung

Da eine Datenanreicherung vor dem eigentlichen Klassifizieren der Fachtexte einen sofortigen Automatisierungseffekt mit sich bringen müsste, wäre es lohnenswert hierzu nähere Untersuchungen durchzuführen. Ein wesentliches Problem besteht darin, Online Fachtexte zu bekommen. Das mag bei der Flut an Dokumenten und Publikationen im Internet zunächst einfach klingen.

Für die Auswahl sind meines Erachtens folgende Kriterien wesentlich:

- die Seriosität der Daten
- die Zugänglichkeit auch der Volltexte
- die rechtlichen Nutzungsmöglichkeiten
- das fachliche Niveau
- die Daten müssen bereits klassifiziert sein, zumindest einer fachlichen Systematik unterliegen

Versucht man eine Datenquelle mit einer Mindestmenge an Dokumenten für die Datenanreicherung zu erhalten, stößt man schnell auf Grenzen. Eine große öffentliche Quelle, die zumindest in Teilen die o. g. Kriterien erfüllt, bietet Wikipedia. Aus diesem Grund habe ich insbesondere das Informatik Portal [Wikipedia: WikiProjekt Informatik/KategorieBaum – Wikipedia] genauer als Quelle zur Datenanreicherung für die CR Classification untersucht.

Da der Wikipedia Kategorienbaum nur wenig Äquivalenz zum CR Classification System hat, müsste zunächst eine intellektuelle Zuordnung der Wikipedia Informatik Kategorien zu den CR Classification Klassen vorgenommen werden.

In einen weiteren Schritt können dann alle Artikeltexte, die diesen Kategorien zugeordnet sind und der Index der einzelnen Klassen der CR Classification als Textkorpus zugeordnet werden. Die Kategorien selbst könnten als Deskriptoren der Klassen genutzt werden.

Für die Zuordnung habe ich zunächst die CR Classification in die Wikipedia eingestellt. Dies ist vielleicht nicht unbedingt notwendig, aber da Wikipedia eine kollaborative

Plattform ist, habe ich mir davon etwas Zuarbeit bei der Zuordnung erhofft. Eine Kritik wurde allerdings sofort an mich gerichtet in Bezug auf: die ausschließlich englische Sprache für die Terme der CR Classification, die Anhäufung von Links und den damit verbundenen mangelnde Informationsgehalt des Artikels. Im Zusammenhang mit dem Erstellen dieses Artikels bin ich von einem ‚Wikipedianer' über ein ähnliches Projekt informiert worden, dem Versuch, Artikel mathematischen Inhalts gemäß der Mathematics Subject Classification zu kategorisieren. Fazit war hier, dass diese Klassifikation für Zeitschriftenartikel und Bücher gedacht ist und für Enzyklopädieartikel ungeeignet ist. Zu viele Begriffe sind in zu vielen Bereichen relevant, eine "tiefe" Kategorisierung wird damit unmöglich, die Kategorien sind überfüllt und unübersichtlich. Die CR Classification ist im gleichen Umfeld angesiedelt, so dass auch hier die gleichen Nachteile zu befürchten sind. Ein weiterer Punkt ist das fachlich sehr unterschiedliche Niveau der Artikel, so dass eine Anreicherung der Klassen zwar vorgenommen werden kann, diese Deskriptoren dann aber nicht den gewünschten Effekt der Automatisierung mit sich bringen, da sie in Fachtexten nicht verwendet werden. Als Beispiel habe ich alle Unterkategorien der Kategorie ‚Objektorientierte Programmierung' als Deskriptoren für die Klasse *D.1.5 Object-oriented Programming* übernommen und anschließend einen Fachtext aus diesem Bereich klassifiziert. Im Ergebnis der Termextraktion kam es zu einer Übereinstimmung mit den Kategorien, die aus Wikipedia übernommen wurden. Der Term *UML* wurde in der Klasse *D.1.5* und im Fachtext gefunden. Insgesamt waren die Terme im Fachtext aber wesentlich spezifischer als die Wikipedia Kategorien zur *Objektorientierten Programmierung*. Meines Erachtens eignen sich mit dem derzeitigen Entwicklungsstand Wikipedia Artikel und deren Kategorien nicht zur Anreicherung einer Fachklassifikation, wie ich sie in meinem Entwurf benötige.

In diesem Zusammenhang ist die Einführung semantisch spezifizierter Verknüpfungen nach dem Semantik-Web-Ansatz, die als Entwicklungsziel für Wikipedia geplant ist, sehr interessant. Dabei sollen Ontologien zur semantischen Beschreibung der Inhalte aufgebaut werden. Wenn es möglich wird, über diese Ontologien Anfragen auszuführen, kann aspektorientiert nach Inhalten recherchiert werden. Das ist im Zusammenhang mit der Suche großer Open Access Datenquellen zu bestimmten Themen wichtig.

Ein weiterer Schwerpunkt zukünftiger Entwicklungen soll die Integration der verschiedenen, bislang selbständigen sprachabhängigen Wikipedia-Versionen sein, mit dem langfristigen Ziel, dem Anspruch an eine Universal Enzyklopädie gerecht zu werden [KUHL05].

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich zunächst mit den Vorraussetzungen und Möglichkeiten, die Klassifikation von Online-Fachtexten aus dem Gebiet der Informatik, zu automatisieren. Das Klassifizieren von Texten in Bibliotheken ist eine Aufgabe des Fachreferenten. Eine Automatisierung kann diese Arbeit erleichtern. In der Arbeit werden zunächst verschiedene Grundlagen zur automatischen Klassifikation beschrieben. Für Fachtexte können andere Strukturen in der Terminologienutzung als für gemeinsprachliche Texte vorausgesetzt werden. Strukturierende Instrumente, die bei der Arbeit des Klassifizierens von Fachtexten zur Anwendung kommen, werden näher untersucht, verglichen und auf ihre Eignung hinsichtlich der Automatisierung bewertet. Aufbauend auf diesen Betrachtungen wurden verschiedenen Methoden ausgewählt und ein Prototyp zur automatischen Klassifikation als Bestandteil einer digitalen Bibliothek entwickelt. Als erster Schritt wird eine Termextraktion vorgenommen, dessen Ziel es ist, die relevanten Terme des Textes zu bestimmen. Die besonderen Eigenschaften von Fachtexten werden bei der Extraktion der relevanten Terme ausgenutzt. Dazu wurde das Verfahren, das am Institut für Informatik, Abteilung automatische Sprachverarbeitung der Universität Leipzig entwickelt wurde, genutzt. Das Open Source Produkt MyCoRe wurde als digitales Content-Repository verwendet. Die Module zur Klassifikation von Dokumenten wurden dahingehend erweitert, dass es möglich ist, Terme den Klassen einer Klassifikation zuzuordnen und über diesen Terme nach Klassen zu recherchieren. Die Module zur Termextraktion wurden an MyCoRe angebunden. Da Texte großen Umfangs auch eine große Menge an Termen als Ergebnis liefern können, wird die Ergebnismenge jeweils für Ein- und Mehrwortterme begrenzt.

Als Fachklassifikation wurde die CR Classification ausgewählt, da sie die im Fachbereich Informatik die am häufigsten angewandte Klassifikation ist. Die XML - Variante der CR Classification wurde in das MyCoRe Klassifikationsmodell transformiert und in die MyCoRe Anwendung integriert. Da die Klassifikation nicht genügend Deskriptoren

gereichert. Für diesen Vorgang wurde eine Nutzeroberfläche implementiert. Sie gestattet es, dem Anwender zum einen jeden Vorschlag der Termextraktion auszuwählen und zum anderen die bereits gefunden Klassifikationen ebenfalls zu übernehmen oder auch abzulehnen. Im Ergebnis liegt ein klassifiziertes Dokument vor und die ausgewählten Terme werden zur Anreicherung der Klassifikation hinzugefügt. Einer Testreihe von 30 bereits intellektuell klassifizierten Dokumenten wurde auf diese Weise klassifiziert und mit den Ergebnissen der manuellen Klassifikation verglichen. Dabei wurde festgestellt, dass die Ergebnisse der Termextraktion im Vergleich zu den manuell vergebenen Schlagworten durchaus akzeptabel sind, eine Automatisierung des Klassifizierens aber aufgrund der zu geringen Termanreicherung noch nicht sichtbar wurde. Insgesamt kann davon ausgegangen werden, das nur dann eine Automatisierung und damit eine Verringerung des Arbeitsaufwandes möglich wird, wenn eine genügend große Menge an Deskriptoren zu den Klassen vorhanden ist. Diese Menge kann entweder durch sukzessive Anreicherung erreicht werden oder aber durch die einmalige Anreicherung der Klassifikation mithilfe weiterer Datenquellen. Dazu wurden einige Kriterien aufgestellt, die bei der Auswahl geeigneter Datenquellen zu beachten sind. Eine interessante Datenquelle könnte Wikipedia sein, in seinem Entwicklungsziel als Semantik Wikipedia. Diese offenen Punkte konnten jedoch in dieser Arbeit nicht mehr beantwortet werden und sollten Gegenstand weiterer Untersuchungen sein.

Literaturverzeichnis

[ACM 98]
ACM Association for Computing Machinery
http://www.acm.org/class/1998/

[BUCH89]
Buchanan, Brian: Bibliothekarische Klassifikationstheorie. - München: Saur, 1989.

[BURK97]
Burkart, Margarete: Thesaurus. In: Buder, M.; Rehfeld, W.; Seeger, T.: Grundlagen der praktischen Information und Dokumentation: Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 4. völlig neu gefasste Ausgabe. - München: KG Saur Verlag, 1997.

[DCMI04]
Dublin Core Metadata Initiative: Dublin Core Metadata Element Set, Version 1.1: Reference Description
http://dublincore.org/

[GEIS99]
Geißelmann, Friedrich: Zur dritten Auflage der RSWK.
In: Bibliotheksdienst, 33. 1999.

[HEYE04]
Heyer, Gerhard: Institut für Informatik, Universität Leipzig
Zusatztexte zur Vorlesung Sprachprodukttechnologie im Sommersemester 2004
http://wortschatz.uni-leipzig.de/asv/lehre/ss04/H-Zusatz-15-Differenzanalyse-280504

[KNOR97]
Knorz, Gerhard: Indexieren, Klassieren, Extrahieren. In: Buder, M.; Rehfeld, W.; Seeger, T.: Grundlagen der praktischen Information und Dokumentation: Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 4. völlig neu gefasste Ausgabe. - München: KG Saur Verlag, 1997.

[KUHL05]
Rainer Kuhlen (2005): Wikipedia – Offene Inhalte im kollaborativen Paradigma – eine Herausforderung auch für Fachinformation
http://www.inf-wiss.uni-konstanz.de/People/RK/Publikationen2005/v5-wikipedia_long-version.pdf

[LORE98]
Lorenz, Bernd: Klassifikatorische Sacherschließung. - Wiesbaden: Harrassowitz, 1998 (=Bibliotheksarbeit. 5).

[MANE04]
Manecke, Hans-Jürgen: Klassifikation, Klassifizieren. In: Kuhlen, Rainer u. a. (Hrsg.): Grundlagen der praktischen Information und Dokumentation. 5. Aufl. - München: Saur, 2004.
Bd. 1. S. 127-140.

[MYCO06]
MyCoRe Starting Guide, Release 1.2, April 2006
F. Lützenkirchen, J. Kupferschmidt, D. Degenhardt u. a.
http://www.mycore.de/cvs/viewcvs.cgi/%7Echeckout%7E/mycore/documentation/StartingGuide/StartingGuide.pdf

[NOHR96]
Nohr, Holger: Systematische Erschließung in deutschen Öffentlichen Bibliotheken. - Wiesbaden: Harrassowitz, 1996 (=Beiträge zum Buch- und Bibliothekswesen. 37).

[POHL2004]
Pohl, Andreas: Text, Datenverteilung in Peer-to-Peer Overlay-Netzwerken
Diplomarbeit 2004, Institut für Informatik, Universität Rostock
http://dbis.informatik.uni-rostock.de/Publikationen/2004.html

[QUAS03]
Quasthoff, Uwe: : Institut für Informatik, Universität Leipzig
Effizientes Dokumentenclustering durch niederfrequente Terme
http://wortschatz.uni-leipzig.de/Papers/Dokumentenclustering.pdf

[RANG69]
Ranganathan, S.R.: Prolegomena to library classification. Bangalore, 1967; zit. n. Nohr

[RSWK98]
Regeln für den Schlagwortkatalog (RSWK) (3., überarbeitete und erweiterte Auflage)
http://deposit.ddb.de/ep/netpub/89/96/96/967969689/_data_stat/www.dbi-berlin.de/dbi_pub/einzelpu/regelw/rswk/rswk_00.htm

[SCHE05]
Scheck, Markus: Automatische Klassifizierung und Visualisierung im Archiv der Süddeutschen Zeitung. In: Medienwirtschaft 1/2005
ASPB Tagung München 2.-7.9. 2005

[SOER69]
Soergel, Dagobert: Klassifikationssysteme und Thesauri – Frankfurt a. M. 1969

[UMST90]
Umstätter, Walter: Wäre es nicht langsam Zeit die Informationstechnologie in der bibliothekarischen Sacherschließung etwas ernster zu nehmen. In: ABI-Technik 11, 1991, Nr. 4

[WITS05]
Witschel, Hans Friedrich: Terminology Extraction and Automatic Indexing Comparison and Qualitative Evaluation of Methods In: Proc. of Terminology and Knowledge Engineering (TKE), 2005
http://wortschatz.uni-leipzig.de/~fwitschel/papers/TKEIndexing.pdf

[WITS051]
Witschel, Hans Friedrich: Text, Wörter, Morpheme - Möglichkeiten einer automatischen Terminologie-Extraktion In: Proc. of GLDV-Tagung 2005
http://wortschatz.uni-leipzig.de/~fwitschel/papers/GLDVPreis.pdf

Anhang

Die im Anhang befindliche CD enthält eine MyCoRe - Webapplikation mit Stand vom: 05.05.2006. Diese Version wurde, um die prototypische Funktionalität der automatischen Klassifikation erweitert. Für die eigenen Erweiterungen ist der Source Code ebenfalls auf der CD.

Dazu gehören:

- Quellen zur Erweiterung des Klassifikationsmodells in MyCoRe (JAVA)
- Quellen zur Anbindung des *Indexers* in MyCoRe (JAVA)
- Quellen für die Nutzerinteraktion, Dokumentauswahl, Termauswahl, Klassifikationszuordnung (JSP, JSTL, JAVA)
- MyCoRe CR Classification (XML - Dokument)
- klassifizierte MyCoRe Metadaten-Dokumentobjekte (XML - Dokument)

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Berlin, den 16.05.2006