

Ein integrierter Ansatz zur wissensbasierten Informationsrecherche

Von der Fakultät Maschinenbau der Universität Stuttgart
zur Erlangung der Würde eines Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Abhandlung

Vorgelegt von

Dipl.-Ing. Christoph Daniel Kunz

aus Friedrichshafen

Hauptberichter: Prof. Dr.-Ing. habil. Prof. e. h. Dr. h. c. mult. Hans-Jörg Bullinger
Mitberichter: Prof. Dr.-Ing. Jürgen Ziegler

Tag der Einreichung: 22. Juni 2005
Tag der mündlichen Prüfung: 16. Dezember 2005

Institut für Arbeitswissenschaft und Technologie-
management (IAT) der Universität Stuttgart

2006

IPA-IAO Forschung und Praxis

Berichte aus dem
Fraunhofer-Institut für Produktionstechnik und
Automatisierung (IPA), Stuttgart,
Fraunhofer-Institut für Arbeitswirtschaft und
Organisation (IAO), Stuttgart,
Institut für Industrielle Fertigung und
Fabrikbetrieb (IFF), Universität Stuttgart
und Institut für Arbeitswissenschaft und
Technologiemanagement (IAT), Universität Stuttgart

Herausgeber:

Univ.-Prof. Dr.-Ing. Prof. E.h. Dr.-Ing. E.h. Dr. h.c. mult. Engelbert Westkämper
und

Univ.-Prof. Dr.-Ing. habil. Prof. E.h. mult. Dr. h.c. mult. Hans-Jörg Bullinger
und

Univ.-Prof. Dr.-Ing. Dieter Spath



I·A·T Institut
Arbeitswissenschaft und
Technologiemanagement
Universität Stuttgart



Fraunhofer Institut
Arbeitswirtschaft und
Organisation

Christoph Daniel Kunz

Ein integrierter Ansatz
zur wissensbasierten
Informationsrecherche

Nr. 436

JUST-JETTER VERLAG
Fachverlag · 71296 Heimsheim

Dr.-Ing. Christoph Daniel Kunz

Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO), Stuttgart

Univ.-Prof. Dr.-Ing. Prof. E.h. Dr.-Ing. E.h. Dr. h.c. mult. Engelbert Westkämper

ord. Professor an der Universität Stuttgart

Fraunhofer-Institut für Produktionstechnik und Automatisierung (IPA), Stuttgart

Univ.-Prof. Dr.-Ing. habil. Prof. E.h. mult. Dr. h.c. mult. Hans-Jörg Bullinger

ord. Professor an der Universität Stuttgart

Präsident der Fraunhofer-Gesellschaft, München

Univ.-Prof. Dr.-Ing. Dieter Spath

ord. Professor an der Universität Stuttgart

Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO), Stuttgart

D 93

ISBN 3-936947-85-6 Jost Jetter Verlag, Heimsheim

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland vom 9. September 1965 in der jeweils gültigen Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtsgesetzes.

© Jost Jetter Verlag, Heimsheim 2006.

Printed in Germany.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Sollte in diesem Werk direkt oder indirekt auf Gesetze, Vorschriften oder Richtlinien (z. B. DIN, VDI, VDE) Bezug genommen oder aus ihnen zitiert worden sein, so kann der Verlag keine Gewähr für die Richtigkeit, Vollständigkeit oder Aktualität übernehmen. Es empfiehlt sich, gegebenenfalls für die eigenen Arbeiten die vollständigen Vorschriften oder Richtlinien in der jeweils gültigen Fassung hinzuzuziehen.

Druck: printsystem GmbH, Heimsheim

Geleitwort der Herausgeber

Über den Erfolg und das Bestehen von Unternehmen in einer marktwirtschaftlichen Ordnung entscheidet letztendlich der Absatzmarkt. Das bedeutet, möglichst frühzeitig absatzmarktorientierte Anforderungen sowie deren Veränderungen zu erkennen und darauf zu reagieren.

Neue Technologien und Werkstoffe ermöglichen neue Produkte und eröffnen neue Märkte. Die neuen Produktions- und Informationstechnologien verwandeln signifikant und nachhaltig unsere industrielle Arbeitswelt. Politische und gesellschaftliche Veränderungen signalisieren und begleiten dabei einen Wertewandel, der auch in unseren Industriebetrieben deutlichen Niederschlag findet.

Die Aufgaben des Produktionsmanagements sind vielfältiger und anspruchsvoller geworden. Die Integration des europäischen Marktes, die Globalisierung vieler Industrien, die zunehmende Innovationsgeschwindigkeit, die Entwicklung zur Freizeitgesellschaft und die übergreifenden ökologischen und sozialen Probleme, zu deren Lösung die Wirtschaft ihren Beitrag leisten muss, erfordern von den Führungskräften erweiterte Perspektiven und Antworten, die über den Fokus traditionellen Produktionsmanagements deutlich hinausgehen.

Neue Formen der Arbeitsorganisation im indirekten und direkten Bereich sind heute schon feste Bestandteile innovativer Unternehmen. Die Entkopplung der Arbeitszeit von der Betriebszeit, integrierte Planungsansätze sowie der Aufbau dezentraler Strukturen sind nur einige der Konzepte, welche die aktuellen Entwicklungsrichtungen kennzeichnen. Erfreulich ist der Trend, immer mehr den Menschen in den Mittelpunkt der Arbeitsgestaltung zu stellen - die traditionell eher technokratisch akzentuierten Ansätze weichen einer stärkeren Human- und Organisationsorientierung. Qualifizierungsprogramme, Training und andere Formen der Mitarbeiterentwicklung gewinnen als Differenzierungsmerkmal und als Zukunftsinvestition in *Human Resources* an strategischer Bedeutung.

Von wissenschaftlicher Seite muss dieses Bemühen durch die Entwicklung von Methoden und Vorgehensweisen zur systematischen Analyse und Verbesserung des Systems Produktionsbetrieb einschließlich der erforderlichen Dienstleistungsfunktionen unterstützt werden. Die Ingenieure sind hier gefordert, in enger Zusammenarbeit mit anderen Disziplinen, z. B. der Informatik, der Wirtschaftswissenschaften und der Arbeitswissenschaft, Lösungen zu erarbeiten, die den veränderten Randbedingungen Rechnung tragen.

Die von den Herausgebern langjährig geleiteten Institute, das

- Fraunhofer-Institut für Produktionstechnik und Automatisierung (IPA),
- Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO),
- Institut für Industrielle Fertigung und Fabrikbetrieb (IFF), Universität Stuttgart,
- Institut für Arbeitswissenschaft und Technologiemanagement (IAT), Universität Stuttgart

arbeiten in grundlegender und angewandter Forschung intensiv an den oben aufgezeigten Entwicklungen mit. Die Ausstattung der Labors und die Qualifikation der Mitarbeiter haben bereits in der Vergangenheit zu Forschungsergebnissen geführt, die für die Praxis von großem Wert waren. Zur Umsetzung gewonnener Erkenntnisse wird die Schriftenreihe „IPA-IAO - Forschung und Praxis“ herausgegeben. Der vorliegende Band setzt diese Reihe fort. Eine Übersicht über bisher erschienene Titel wird am Schluss dieses Buches gegeben.

Dem Verfasser sei für die geleistete Arbeit gedankt, dem Jost Jetter Verlag für die Aufnahme dieser Schriftenreihe in seine Angebotspalette und der Druckerei für saubere und zügige Ausführung. Möge das Buch von der Fachwelt gut aufgenommen werden.

Engelbert Westkämper Hans-Jörg Bullinger Dieter Spath

Vorwort

"Als ich des Suchens müde war, erlernte ich das Finden." (Friedrich Nietzsche)

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Arbeitswissenschaft und Technologiemanagement (IAT) der Universität Stuttgart und in enger Zusammenarbeit mit dem Fraunhofer Institut für Arbeitswirtschaft und Organisation (IAO). Mein Interesse an der hier bearbeiteten Fragestellung wurde hauptsächlich innerhalb zweier, vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Forschungsprojekte geweckt, dem Leitprojekt INVITE - Intuitive Mensch-Technik Interaktion für die vernetzte Informationswelt der Zukunft - und AWAKE Networked Awareness for Knowledge Discovery. In ihnen entstanden die grundlegenden Ideen, wesentliche Ergebnisse, sowie eine erste prototypische Umsetzung. Seit den frühen 1970er Jahren gewinnen Suche und Recherche nicht zuletzt aufgrund des Durchbruchs des Internets zunehmend an Bedeutung. Es erschien mir daher lohnenswert, sich mit dieser Thematik im Rahmen einer Dissertation auseinander zu setzen.

Meinem Doktorvater Herrn Prof. Dr.-Ing. Hans-Jörg Bullinger, Präsident der Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V., danke ich besonders für seine wohlwollende Unterstützung und Förderung meiner Arbeit, sowie für die Übernahme des Hauptberichts.

Besonderer Dank gilt dem Mitberichter Herrn Prof. Jürgen Ziegler, meinem ehemaligen Abteilungsleiter am Fraunhofer IAO und nun Professor für Interaktive Systeme an der Universität Duisburg für die Mühen bei der inhaltlichen Begleitung, konstruktive Diskussion und Förderung meiner Arbeiten und für die richtungweisenden Impulse zur rechten Zeit.

Ebenfalls möchte ich mich bei allen Kollegen und wissenschaftlichen Hilfskräften des Competence Centers Human-Computer Interaction insbesondere bei Prof. Dr. Frank Heidmann, mein ebenfalls ehemaliger Abteilungsleiter und nun Professor für Interfacedesign an der Fachhochschule Potsdam, für die hilfreichen Anregungen und wissenschaftlichen Gespräche sowie für die fruchtbare Arbeitsatmosphäre bedanken.

Diese Arbeit wäre nicht möglich gewesen ohne die großzügige und geduldige Unterstützung meiner Eltern, Wunnibald und Christine Kunz, die mir so vieles ermöglicht haben. Dafür will ich an dieser Stelle herzlichst danken. Nicht zuletzt gilt großer Dank meiner Freundin Dr. Patricia Wolf für ihre große Geduld, ihr Verständnis und ihre beständige liebevolle Unterstützung während der Erstellung dieser Arbeit und darüber hinaus. Ihr und meinem langjährigen Freund Holger Oehme danke ich für beständige Motivation sowie Rat und Tat in allen Lebenslagen.

Stuttgart, Januar 2006

Christoph Daniel Kunz

Inhaltsverzeichnis

1	Einleitung	16
1.1	Problemstellung	17
1.2	Zielsetzung	18
1.3	Vorgehensweise	20
2	Methoden des Information Retrieval	22
2.1	Begriffsklärung, Grundlagen und Vorannahmen	22
2.1.1	Datum, Information und Wissen	22
2.1.2	Vermittlung von Information durch Kommunikation	24
2.1.3	Soziotechnische Systeme	25
2.2	Einführung in das Gebiet des Information Retrieval	27
2.2.1	Qualitätsbeurteilung	28
2.2.2	Komponenten eines Information Retrieval Systems	29
2.3	Manuelle Erschließungsverfahren	30
2.3.1	Klassifikationen	31
2.3.2	Thesauern	33
2.4	Automatische Erschließung und Retrieval Modelle	34
2.4.1	Klassifikation von IR-Modellen	35
2.4.2	Termgewichtung	36
2.4.3	Boolesches Retrieval	37
2.4.4	Vektorraummodell	38
2.4.5	Probabilistische Modelle	39
2.4.6	Diskussion	40
2.5	Wissensbasiertes Information Retrieval	41
2.5.1	Ontologien und semantische Netze	42
2.5.1.1	Aussagelogiken	44
2.5.1.2	Beschreibungslogiken	45
2.5.1.3	Ontologiesprachen im weltweiten Netz	47
2.5.1.4	Methoden und Verfahren zur Generierung von Ontologien	49
2.5.2	Verarbeitung der natürlichen Sprache	52
2.5.3	Logische Ansätze	54
2.5.4	Diskussion	55
2.6	Informationsvisualisierung im Information Retrieval	57
2.6.1	Visualisierungen von vernetzten Begriffssystemen	59
2.6.2	Visualisierungen zur Unterstützung der Anfrageformulierung	60
2.6.3	Visualisierungen zur Kontextualisierung von Suchergebnissen	62
2.6.4	Diskussion	64
3	Ein integrierter Ansatz zur wissensbasierten Informationsrecherche	68
3.1	Anforderungen an einen integrierten Ansatz zur Informationsrecherche	68
3.2	Grundlegender Ansatz und Begründung	69
3.2.1	Argumentation des Ansatzes	71
3.3	Semantisches Modell der ontologischen Wissensbasis	73
3.3.1	Anwendung und Anforderungen	74
3.3.1.1	Anwendung	74
3.3.1.2	Anforderungen	75
3.3.2	Spezifikationssprache einer probabilistischen Ontologie	77
3.3.2.1	OWL Full	77
3.3.2.2	pOWL - eine probabilistische Erweiterung von OWL	83
3.3.2.3	Einbettung der Wahrscheinlichkeiten in OWL mittels Reifikation	85
3.3.3	Minimalabstraktionen	86

3.3.4	Integriertes Suchmodell	89
3.3.4.1	Stichwortsuche	90
3.3.4.2	Semantische Suche	90
3.4	Ein kollaboratives Verfahren zur Wissensakquise	90
3.4.1	Grundlegende Hypothesen und Argumentation	92
3.4.2	Aufgabenstellung	94
3.4.3	Kollaboratives Indexieren	94
3.4.3.1	Ablagestrukturen	95
3.4.3.2	Behandlung von geordneten Lesezeichen	95
3.4.3.3	Integration lose abgelegter Lesezeichen in die Ontologie	98
3.5	Schnittstelle zur Informationssuche	98
3.5.1	Visualisierung und Navigation ontologischer Information - der MatrixBrowser	99
3.5.1.1	Grundidee: Matrixvisualisierung	100
3.5.1.2	Interaktives Verhalten	101
3.5.1.3	Extraktion von Teilhierarchien	103
3.5.2	Visuelle Konstruktion semantischer Suchanfragen	104
3.5.2.1	Einschränkung	104
3.5.2.2	Visuelle Abbildung	105
3.5.3	Integration der Techniken	106
3.5.3.1	Aufbereitung der Ontologie	107
3.5.3.2	Kopplung von Matrix- und Tabellendarstellung	107
3.5.3.3	Kontextualisierung von Suchergebnissen	107
3.5.3.4	Dreistufiger Rechercheprozess	108
4	Evaluierung	109
4.1	Evaluierung des Verfahrens zur kollaborativen Wissensakquise	109
4.1.1	Ziel und Fragestellung	109
4.1.2	Vorgehensweise	110
4.1.3	Ergebnisse	112
4.1.3.1	Browserlesezeichen für einen offenen Informationsraum	112
4.1.3.2	Wissenskarten für einen geschlossenen Informationsraum	114
4.1.4	Diskussion	115
4.2	Iterativer Gestaltungsprozess bei der Entwicklung des MatrixBrowsers	116
4.2.1	Erste Iteration	117
4.2.1.1	Vorgehensweise	117
4.2.1.2	Ergebnisse	118
4.2.1.3	Diskussion	119
4.2.2	Zweite Iteration	119
4.2.2.1	Vorgehensweise	119
4.2.2.2	Ergebnisse	120
4.2.2.3	Diskussion	120
4.2.3	Dritte Iteration	120
4.2.3.1	Vorgehensweise	121
4.2.3.2	Ergebnisse	121
4.2.3.3	Diskussion	121
4.3	Evaluierung der Rechterschnittstelle	122
4.3.1	Ziel und Fragestellung	122
4.3.2	Vorgehensweise	122
4.3.3	Ergebnisse	123
4.3.4	Diskussion	125
5	Zusammenfassung und Ausblick	127
	Literaturverzeichnis	129

Abbildungsverzeichnis

Abbildung 1.1:	Rahmenwerk zur wissensbasierten Informationsrecherche	19
Abbildung 1.2:	Aufbau der Arbeit	21
Abbildung 2.1:	Daten - Information - Wissen	23
Abbildung 2.2:	Grundschema des Information Retrievals	29
Abbildung 2.3:	Klassifikation von Retrievalmodellen	34
Abbildung 2.4:	Ontologiesprachen auf Basis von RDF	47
Abbildung 2.5:	Beispielontologie in RDFS	48
Abbildung 2.6:	Rahmenwerk zur semiautomatischen Erstellung von Ontologien	51
Abbildung 2.7:	Referenzmodell der Informationsvisualisierung	57
Abbildung 2.8:	Verzerrungsbasierte Visualisierungstechniken mit Fischaugeneffekt	59
Abbildung 2.9:	3D-Visualisierung von Themenkarten	60
Abbildung 2.10:	Formulierung semantischer Suchanfragen	62
Abbildung 2.11:	Visualisierungen zur Erklärung der Ergebnismenge	63
Abbildung 2.12:	Visualisierungen zur Kontextualisierung von Suchergebnissen	63
Abbildung 3.1:	Das Suchproblem	69
Abbildung 3.2:	Integrierter Ansatz zur wissensbasierten Informationsrecherche	70
Abbildung 3.3:	Analytisch Klassifikation vs. Facettenklassifikation	75
Abbildung 3.4:	Beispiel einer Reifikation	76
Abbildung 3.5:	Klassenbeschreibung in OWL	78
Abbildung 3.6:	Minimal benötigte Abstraktionen	86
Abbildung 3.7:	Modellierungsbeispiel einer Wissensdatenbank	89
Abbildung 3.8:	Grundlegende Idee der kollaborativen Wissensakquise	91
Abbildung 3.9:	Beispielhafter Ablauf des kollaborativen Indexierens	97
Abbildung 3.10:	Prototyp des MatrixBrowsers	100
Abbildung 3.11:	Strategien zur Extraktion von Teilhierarchien	104
Abbildung 3.12:	Anfrageleiste zur Konstruktion semantischer Abfragen	105
Abbildung 3.13:	Schnittstelle zur Informationssuche	106
Abbildung 4.1:	Statistiken von Begriffsbeschreibungen nach der Auswertung von Browserlesezeichen	113
Abbildung 4.2:	Statistiken von Begriffsbeschreibungen nach der Auswertung von Wissenskarten	114
Abbildung 4.3:	Benutzerorientierter Gestaltungsprozess nach ISO 13407	116
Abbildung 4.4:	Visuelle Suchpfade von Probanden	118
Abbildung 4.5:	Gestaltungsvarianten des MatrixBrowsers	121
Abbildung 4.6:	Ergebnisse der Evaluationsstudie der Recherveschnittstelle	124

Tabellenverzeichnis

Tabelle 2.1:	Daten- vs. Information Retrieval	28
Tabelle 2.2:	Beispiel "Alle Menschen sind sterblich" für Aussagenlogiken	45
Tabelle 2.3:	Beispiel "Alle Menschen sind sterblich" für Beschreibungslogiken	46
Tabelle 3.1:	Syntax und Semantik von Begriffsbeschreibungen.	81
Tabelle 3.2:	Erfüllbarkeitsbedingungen von Interpretationen	82
Tabelle 3.3:	Menge der Dublin Core Elemente	88

Zusammenfassung

Dokumente im Internet als auch in betrieblichen Kontexten (Intranet, Portal, Wissensdatenbank) liegen selten in strukturierter Form vor und besitzen fast keine Metainformationen bezüglich ihres Inhalts. Schon die schiere Menge, aber auch das Nichtvorhandensein einer einheitlichen Zugangsstruktur erschwert sowohl deren menschliche als auch deren maschinelle Verarbeitung und Wahrnehmung, im Besonderen das Entdecken relevanter Information, deren Zusammenhang und deren Synthese zu einem umfassenden Gesamtbild.

Gerade in Unternehmen des Wirtschaftssystems wird es inzwischen als unabdingbar angesehen, durch einen schnellen und reibungslosen Informationszugriff eine Erhöhung organisationaler Lern- und Wandlungsfähigkeit zu fördern, um dadurch auf eine immer turbulenter werdende Umwelt reagieren zu können¹. Um dieses zu erreichen, werden Instrumente wie Intranets oder Wissensdatenbanken im Rahmen von Expertengemeinschaften² eingesetzt, um vorhandenes (bereits kodifiziertes) Wissen zu konservieren und zugänglich zu machen. Dabei sind effiziente und nutzerfreundliche Zugriffs- und Recherchemechanismen von besonderer Bedeutung, um eine schnelle Lokalisierung von gerade wichtiger Information zu ermöglichen.

Auf Grundlage dieser Ausgangssituation wird in der vorliegenden Arbeit ein nutzerorientierter und ganzheitlicher Ansatz beschrieben und evaluiert, welcher eine semantische Erschließ- und Recherchierbarkeit von großen und vernetzten Informationsräumen unter Einbeziehung der menschlichen Nutzung und Interpretation ermöglicht. Unter dem Begriff Informationsraum werden außer Dokumentkollektionen auch sonstige Ressourcen zusammengefasst, deren Inhalt von Interesse ist (Dienste, Produktdaten, etc.). Den Anwendungskontext stellen Intranets und komplexe Portalseiten, sowie spezialisierte Dokumentkollektionen dar, welche von Expertengruppen zur Dokumentation gemeinsamer Erfahrungen erstellt werden. In diesen Anwendungsgebieten ist ein effektiver und nutzerfreundlicher Informationszugriff essentiell.

Der Ansatz verbindet eine herkömmliche stichwortbasierte Volltextsuche mit einer semantischen Suche auf Grundlage einer lernenden Themenontologie, welche einen Informationsraum abstrahiert und verdichtet. Ausgangspunkt ist dabei die Annahme, dass in Situationen der Informationssuche ein Kommunikationsprozess zwischen dem menschlichen Nutzer und dem benutzen Werkzeug statt findet. Neue ontologische Begriffe werden dabei durch Beobachtung der in diesem Kommunikationsprozess entstandenen Artefakte erzeugt, wie explizierte Präferenzen und persönliche Klassifikationssysteme, die sich in Lesezeichen zur Merkhilfe manifestieren. Eine Ontologie als Modellierungsformalismus ermöglicht zudem die einfache Verknüpfungsmöglichkeiten dieser mit nicht textuellen, strukturierten Informationsquellen (z.B. Datenbankschemata). Diese können dadurch parallel mit stichwortbasierten Suchanfragen ausgewertet und gemeinsam als Antworten dargestellt werden³.

Die verschiedenen damit geschaffenen Arten des Informationszugriffs erfolgen über eine einheitliche und bedarfsgerechte Nutzerschnittstelle. Deren Hauptmerkmal ist die Visualisierung der Ontologie als Abstraktion und Verdichtung der semantischen Struktur des suchbaren Informationsraums. Die dabei eingesetzte Technik beruht auf der hochinteraktiven Darstellung einer Adjazenzmatrix des Ontologiegraphen. Diese Strukturvisualisierung ist mit einer Präsentation von Suchergebnissen zu einem wählbaren Thema eng gekoppelt. Zusätzlich sind damit stichwortbasierte und semantische Abfragemöglichkeiten verbunden. Semantische Suchanfragen können innerhalb der Nutzerschnittstelle unter Verwendung der dargestellten Ontologie visuell konstruiert werden, womit die Erlernung einer logischen Abfragesprache (z.B. F-Logik⁴) entfällt. Damit werden drei Arten von Informationszugängen geschaffen: Auf Strukturebene kann navigatorisch durch Verbreiterung und Einengung eines Themas, sowie der Verfolgung von bedeutungsvollen Querbeziehungen gestöbert und unmittelbar die zugehörigen Inhalte eingesehen und recherchiert werden. Eine Volltextsuche ermöglicht eine ungenaue Suche auf Basis von Stichwörtern, während die semantische Suche eine präzise Lokalisierung von gewünschter Information ermög-

¹vgl. WOLF: *Erfolgsmessung der Einführung von Wissensmanagement*, 2003, S. 1

²WENGER: *Communities of Practice: Learning, Meaning, and Identity*, 1998

³vgl. STUDER/SCHNURR/NIERLICH: *Semantisches Knowledge Retrieval*, 2001, S. 14

⁴KIFER/LAUSEN/WU: *Logical Foundations of Object-Oriented and Frame-Based Languages*, 1995

licht. Mit Hilfe dieser Mechanismen soll auch das Verständnis des eigenen Informationsbedarfs der Nutzer gefördert werden, welcher das Ziel einer Recherche bestimmt.

Eine umfangreiche Evaluation des Ansatzes zeigt Leistungsvorteile gegenüber der herkömmlichen Volltextsuche und belegt dessen Nützlichkeit in den Anwendungsbereichen. Es konnte gezeigt werden, dass die prototypische Umsetzung des Ansatzes gleichermaßen von Experten und durchschnittlichen Nutzern zu bedienen war. Ebenfalls konnte die Machbarkeit der kollaborativen Wissensakquise durch Beobachtung des Ablageverhaltens und die hohe Qualität der daraus entstandenen Wissensbasis nachgewiesen werden. Neben der Güte der dadurch erzeugten neuen Begrifflichkeiten ist die damit erreichte Klassifikationsquote hervor zu heben, welche weit über rein maschinellen Verfahren liegt.

Summary

Documents on the internet as well as in business contexts (intranet, portals, knowledge database) are rarely available in a structured format and hold hardly any metainformation regarding the content. Sheer quantity as well as the non-availability of a standardized access structure complicates human and machine processing and perception, i.e., finding of relevant information, their context and their synthesis into a comprehensive overall picture.

Especially in businesses of the economic system a fast and unobstructed information access is considered indispensable for facilitating organizational mutability and ability to learn in order to be able to react to an environment being more and more turbulent⁵. For this reason, instruments such as intranets and knowledge databases are deployed for conserving and reusing existing (already codified) knowledge in the context of Communities of Practice (CoPs⁶). Besides, efficient and user-friendly information access and retrieval mechanisms are of great importance for enabling rapid localization of just now relevant information.

On the basis of this initial situation, the present work describes and evaluates a user-centered and holistic approach for making large and networked information spaces semantically available and researchable by including human use and interpretation. The term information room includes not only document collections but also other resources whose content is of interest (services, product data, etc.). Application areas are intranets and complex portal sites as well as specialised document collections which are created by expert communities in order to preserve their common experience. An effective and user-friendly information access is crucial in these application areas.

The approach combines conventional keyword-based full-text search with semantic retrieval methods on the basis of a learning probabilistic ontology of topics which abstracts and condenses an information space. The starting point is the assumption that in the case of a research situation a communication process takes place between a human user and a used tool and forms a socio-technical system between the communication partners. New ontological concepts are generated by observing artefacts which are created in this communication process and can be explicit preferences and personal classification systems that become manifest in bookmarks as a memory aid. Using an ontology as a modelling formalism allows a simple linking-up with non-textual and structured data sources (e.g. data-base schemes). These can be then analysed simultaneously with keyword-based queries and displayed as answers at the same time⁷.

The various resulting information access mechanisms are integrated into one consistent and requirements-based user interface. The main feature of the user interface is the visualization of the ontology, abstracting and aggregating the semantic structure of the researchable information space. For this, the technique applied is based on a highly interactive adjacency matrix display of the ontology graph. The structure visualization is tightly coupled with a list-based presentation of search results given by a freely selectable topic based on the principle of direct manipulation. Additionally, keyword-based and semantic query features are combined with it. The semantic queries can be visually constructed in order to avoid learning of a logical query language, like F-Logic⁸. Thus, three types of an information access are created: browsing on a structural level can be done by navigationally broadening or narrowing a topic and by following semantic relationships between topics. Corresponding content can be seen directly and inspected in the list. A full-text search on the basis of keywords provides an imprecise search, whereas by the use of semantic queries, needed information can be precisely localized. By using these mechanisms, the understanding of users' information needs should be furthered which directs the goal of a research.

Extensive evaluation studies of the approach show performance advantages in comparison with a conventional full-text retrieval and attests usefulness in the application areas. It could be demonstrated that the prototypical implementation of the approach could be used by experts and

⁵cp. WOLF: *Erfolgsmessung der Einführung von Wissensmanagement*, 2003, S. 1

⁶WENGER: *Communities of Practice: Learning, Meaning, and Identity*, 1998

⁷cp. STUDER/SCHNURR/NIERLICH: *Semantisches Knowledge Retrieval*, 2001, S. 14

⁸KIFER/LAUSEN/WU: *Logical Foundations of Object-Oriented and Frame-Based Languages*, 1995

customary users alike. Also, the feasibility of the collaborative knowledge acquisition by observing users' filling behaviour as well as the high quality of the resulting knowledge base could be proven. Besides, the quality of the new concepts generated, the exceedingly high classification quota has to be stressed which is a more advanced performance compared to pure mechanical classification methods.

1 Einleitung

Vor der Entstehung der Schrift war das persönliche Gedächtnis die einzige Möglichkeit, Erfahrungen und Wissen zu konservieren. Auch existierte nur eine einzige Möglichkeit, sein Wissen an andere Menschen weiterzugeben. Entweder musste es ihnen persönlich erläutert oder aber durch eine dritte Person ausgerichtet werden. Um diese Einschränkungen zu überwinden, entwickelten die Sumerer, eine Hochkultur im heutigen Irak, schon 3000 v. Chr. die ersten Schriftzeichen. Die Schrift ermöglicht bis heute nicht nur die Kommunikation über zeitliche und räumliche Grenzen hinweg, sondern bildet auch die Grundlage von Tradition, Kultur und Bildung durch die mittelbare Weitergabe von Wissen.

Obwohl gerade letzterer Sachverhalt aus heutiger Sicht leicht nachzuvollziehen ist, entstanden die ersten Bibliotheken und Archive aus der Motivation heraus, eine öffentliche Verwaltung (Bürokratie) überhaupt erst zu ermöglichen. Mit dem Kopieren und Speichern von Texten aber geht die Notwendigkeit einher, diese wieder aufzufinden. Mit zunehmender Menge wurde es daher immer wichtiger, abgelegte Inhalte systematisch zu ordnen, damit sie wieder aufgefunden werden konnten. Die erste nachgewiesene Bibliothek, welche das Kennzeichen der planvollen Sammlung¹ besitzt, ist die Bibliothek Assurbanipals (668-627 v. Chr.) in dessen Residenz in Ninive. Als Erschließungssystem wurden Etikettentäfelchen benutzte, die an Tafelbehältnisse angelehnt wurden und die den Inhalt der Werke thematisch umschrieben. Typische Bezeichnungen waren zum Beispiel "Inspektionen", "Rechnungsabschlüsse" oder auch "Herden und Hirten" und "Tempel und Niederlassungen". Es ergibt sich ein, wenn auch einfaches, Klassifikationssystem.

Die Bibliothek von Alexandria (Museion) und ihre "Schwester", jene von Pergamon, sind die herausragenden Bibliotheken der griechisch-römischen Antike. Die alexandrinische Bibliothek ist die mit Abstand größte der Epoche und gilt als das antike Vorläufermodell der modernen Nationalbibliotheken. Von Bedeutung ist sie besonders weil in ihr die ersten Bibliothekskataloge und antike Vorbilder moderner Buchverzeichnungen geschaffen wurden. Über die Anzahl der aufbewahrten Schriftrollen, die im Laufe der 550jährigen Geschichte der Bibliothek von Alexandria gesammelt wurden, herrscht keine Klarheit. Zu divergierend sind die in den Quellen genannten Zahlen, die von 40'000 bis 700'000 Rollen reichen, welche es zu verwalten galt. Der erste schriftliche Katalog der Bibliotheksgeschichte überhaupt stammt von Kallimachos von Kyrene (305-240 v. Chr.)². Es handelt sich zwar noch nicht um einen Bibliothekskatalog im modernen Sinne, denn die auf 120 Rollen geschriebenen, so genannten Pinakes, waren nicht für die Nutzer bestimmt und verzeichneten nicht den Gesamtbestand der Bibliothek, sondern konzentrierten sich auf eine Auswahl der griechischen Schriftsteller. Kallimachos sortierte die Titel nach wissenschaftlichen und literarischen Kategorien gemäß aristotelischem Vorbild.

Seit dem Bau des ersten Computers 1958 auf Basis von Transistorentechnik und dem Beginn des Internets 1983 als das TCP/IP³ Kommunikationsprotokoll sich zum Standard des ARPANETs⁴ etablierte, konnten große Informationsmengen nicht nur auf kleinstem Raum untergebracht, sondern auch schnell darauf zugegriffen werden. Seitdem hat sich die Wissenskultur in der modernen, postindustriellen Informationsgesellschaft nachhaltig verändert und digitale Online-Archive gewinnen als zentrale gesellschaftliche Instanz immer mehr an Bedeutung. Die Suche nach Information zu einem bestimmten Thema hat sich von einer Aufgabe in (wissenschaftlichen) Bibliotheken und Sammlungen zu einer alltäglichen Angelegenheit vieler Menschen entwickelt. Im Netz gefunden zu werden ist nicht nur für Unternehmen ein entscheidender Erfolgsfaktor. Es erstaunt daher nicht, dass in verschiedensten Wissenschaften Themen wie die Verwaltung von Wissen, sowie organisationales Lernen und Gedächtnis Konjunktur haben.

Gerade in Unternehmen des Wirtschaftssystems wird es inzwischen als unabdingbar angesehen, durch einen schnellen und reibungslosen Informationszugriff die organisationale Lern- und Wandlungsfähigkeit zu fördern, um dadurch auf eine immer turbulenter werdende Umwelt reagie-

¹ OPPENHEIM: *Ancient Mesopotamia: Portrait of a Dead Civilization*, 1964

² BLUM: *Kallimachos und die Literaturverzeichnung bei den Griechen*, 1977

³ Transmission Control Protocol und Internet Protocol

⁴ Netzwerk der Advanced Research Projects Agency, auch DARPA, einer an das US-Verteidigungsministerium angeschlossene Forschungseinrichtung

ren zu können⁵. Dafür werden Instrumente wie Intranets oder Wissensdatenbanken eingesetzt, um vorhandenes (bereits kodifiziertes) Wissen zu konservieren und zugänglich zu machen. Hierzu sind effiziente und nutzerfreundliche Zugriffs- und Rechercheinstrumente von besonderer Bedeutung, welche eine schnelle Lokalisierung von gerade wichtiger Information ermöglichen.

"Information Retrieval als wissenschaftliche Disziplin, die die inhaltliche Suche nach Informationen in Sammlungen von Dokumenten untersucht und Modelle, Methoden und Verfahren dafür entwickelt, hat dadurch aber nicht entsprechend größere Beachtung gefunden. Häufig werden [...] eher einzelne Technologien und Dienste wahrgenommen als eine zusammenfassende Sicht aus der Perspektive der inhaltlichen Suche."⁶

1.1 Problemstellung

Dokumente im Internet als auch Informationsräume in betrieblichen Kontexten (Intranet, Wissensdatenbank) liegen selten in strukturierter Form vor und besitzen fast keine Metadaten (Daten über Daten) bezüglich ihres Inhalts. Schon die schiere Menge, aber auch das Nichtvorhandensein einer einheitlichen Zugangsstruktur erschwert sowohl deren menschliche als auch deren maschinelle Verarbeitung und Wahrnehmung, das heißt das Entdecken von Information, deren Zusammenhang und deren Synthese zu einem umfassenden Gesamtbild.

Im Wesentlichen existieren drei Ansätze, um umfangreiche, vernetzte Informationsräume erschließbar zu machen und somit Rechercheprozesse zu unterstützen:

1. Für den menschlichen Zugang zu Informationsräumen existiert der aus dem Bibliothekswesen bekannte Ansatz der Systematik, in dem Inhalte nach Themengebieten hierarchisch katalogisiert und nach Autoren, Erscheinungsjahr, etc. sortiert werden. Im übertragenen Sinn und stark vereinfacht folgen auch Portalsysteme⁷ diesem Ansatz. Der Nutzer eines solchen Zugangs erhält damit eine navigierbare Struktur mit der er nach dem Prinzip des kontinuierlichen Vordringens bis auf die gewünschte Ressourcenebene vordringen kann. Dies setzt aber hochgradige redaktionelle Vorarbeiten der Systematik voraus. Des Weiteren sind die damit gebotenen Zugangsmechanismen, wie rollen- und aufgabenbasierte oder hierarchische Kategoriensysteme, zu starr, um flexible und individuelle Sichten auf einzelne Inhalte zu bieten und somit ein Gesamtbild über eine Menge von Inhalten zu vermitteln.
2. Auf Seiten der maschinellen Verarbeitung hat die herkömmliche Stichwortsuche die weiteste Verbreitung gefunden, obwohl sie der vernetzten Natur von Informationsräumen nur wenig Rechnung trägt. Damit lässt sich fast keine Information über die Struktur und Beziehungen der Inhalte untereinander gewinnen. Auch fokussieren die meisten dieser Suchtechniken auf mathematische Modelle und Algorithmen für die Lokalisierung von Information. Dadurch sind aus Sicht der Nutzer viele solcher Systeme schwer zu nutzen, da sich Informationsbedürfnisse schwer formulieren und schwierig in Systemanfragen ausdrücken lassen⁸. Nicht ungewöhnlich ist dabei die Eingabe von Suchtermen, welche mit den Indextermen der Retrievalsysteme nicht übereinstimmen und daher keine Ergebnisse liefern. Auf zu unspezifische Suchterme reagieren sogar modernste Systeme wie *Google* mit Suchergebnissen im Millionenumfang.
3. Als neuerer Ansatz erhält die Vision des *Semantischen Netzes*⁹ derzeit viel Beachtung. Dabei ist der allgemeine Grundgedanke die Anreicherung der dargebotenen Medien des bestehenden Internets mit Metadaten, so dass deren inhaltliche Bedeutung (Semantik) von

⁵vgl. WOLF: *Erfolgsmessung der Einführung von Wissensmanagement*, 2003, S. 1

⁶FERBER: *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*, 2003, Vorwort

⁷z.B. GURZKI/HINDERER/EBERHARD: *Marktübersicht Portal Software - Marktübersicht für Business-, Enterprise-Portale und E-Collaboration*, 2002

⁸vgl. BELKIN: *Anomalous States of Knowledge as a Basis for Information Retrieval*, 1980

⁹z.B. BERNERS-LEE: *Semantic Web Roadmap*, 1998

Automaten versteh- und interpretierbar wird. Bezogen auf die Informationsrecherche werden hierbei Inhalte und/oder ganze Informationsräume mittels komplexer Wissensstrukturen wie semantischen Netzen, Ontologien oder Frames repräsentiert. Ziel dieser, aus dem Gebiet der Künstlichen Intelligenz stammenden Verfahren ist es, Wissen unabhängig von der sprachlichen "Oberflächenstruktur" zu modellieren. Aufgrund diesen Wissensmodellen können dann automatische Schlussfolgerungen abgeleitet werden. Dies eröffnet eine Vielzahl neuer Möglichkeiten bei der maschinellen Verarbeitung von Inhalten. Da die Maschine nicht nur ihre Semantik erfasst, sondern vielleicht auch weiß, dass dieselbe Informationseinheit an anderer Stelle, aber in anderer Form, schon vorhanden ist, kann Information leichter aggregiert, integriert und wieder verwendet, aber vor allem leichter gefunden werden. Gerade Wortambiguitäten, das Hauptproblem der stichwortbasierten Suche, können durch Einbeziehung von Kontexten aufgelöst werden.

Suchmethoden auf Basis von Ontologien versprechen zwar einige Vorteile gegenüber merkmalsbasierten Ansätzen, welche Wörter als bloße Zeichenketten auffassen. Nachteile ergeben sich jedoch aus der Aufwändigkeit der Erstellung der zugrunde liegenden Wissensbasis. Diese muss für eine automatische Schlussfolgerung widerspruchsfrei sein und kann aus diesem Grund bisher nur schwer automatisiert gebildet werden. Zudem muss die zugrunde liegende Informationsarchitektur eine hohe Qualität der Struktur und eine gemeinsame Akzeptanz der Nutzer aufweisen. Die Erstellung erfordert daher einen erheblichen redaktionellen Aufwand. Beim direkten Abfragen der Wissensbasis können nur exakte Treffer erzielt werden, da die Suchterme mit der ontologischen Terminologie abgeglichen werden. Nur wenn die Suchbegriffe in der Wissensbasis enthalten sind, werden Antworten deduziert. Zudem erfolgt keine Sortierung der Ergebnisinformation nach Relevanz, da Ähnlichkeiten und Klassenzugehörigkeiten aufgrund der meistens verwendeten booleschen Wahrheitsfunktion nicht gewichtet werden können. Begriffe der Ontologie werden daher meist nur zur Verbreiterung einer Suchanfrage genutzt, die an eine herkömmliche Stichwortsuchmaschine weitergeleitet wird.

Ein weiteres Problem ergibt sich aus der Komplexität semantischer Ansätze, nämlich das der Gestaltung geeigneter Nutzerschnittstellen, welche die eigentlichen Vorteile erst nutzbar machen und den Informationszugriff ermöglichen. Gestaltungsfragen betreffen die Darstellung der in einer Ontologie formalisierten Informationsstruktur, sowie einfache Interaktions- und Visualisierungstechniken zur Konstruktion logischer Abfragen. Es erstaunt, dass erst in neuerer Zeit die Informationsvisualisierung auch im Information Retrieval Aufmerksamkeit findet. Dies kommt erstmals 1999 im "Aufruf zur Mitgestaltung" der ACM SIGIR, eine der führenden Information Retrieval Konferenzen, zum Ausdruck:

"In der Information Retrieval-Gemeinschaft verfestigt sich die Meinung, dass ein Schlüssel zur Verbesserung von Informationszugriffssystemen in der stärkeren Fokussierung auf die Mensch-Computer Schnittstelle liegt"¹⁰.

1.2 Zielsetzung

Ziel der vorliegenden Arbeit ist die Entwicklung und Evaluierung eines nutzerorientierten und ganzheitlichen Ansatzes für ein interaktives Computersystem, welcher es erlaubt, große und vernetzte Informationsräume unter Einbeziehung der menschlichen Nutzung und Interpretation semantisch erschließ- und recherchierbar zu machen. Den Anwendungskontext stellen Intranets und komplexe Portalseiten, sowie spezialisierte Dokumentkollektionen dar, welche von Expertengruppen zur Dokumentation gemeinsamer Erfahrungen erstellt werden¹¹.

Die nutzerorientierte Gestaltung ist eine Art der Entwicklung interaktiver Systeme, die sich darauf konzentriert, Systeme gebrauchstauglich zu machen. Die Anwendung des Wissens über

¹⁰ übersetzt aus HEARST: *SIGIR: Call For Participation*, 1999

¹¹ z.B. im Rahmen sogenannter Communities of Practice (CoPs), siehe u.a. WENGER: *Communities of Practice: Learning, Meaning, and Identity*, 1998

menschliche Faktoren und Ergonomie bei der Gestaltung interaktiver Systeme erhöht deren Effektivität und Effizienz, verbessert die Arbeitsbedingungen des Menschen und wirkt möglichen nachteiligen Auswirkungen auf die menschliche Gesundheit, Sicherheit und Leistung entgegen. Wird die Ergonomie bei der Gestaltung von Systemen angewandt, sind menschliche Fähigkeiten, Fertigkeiten und Bedürfnisse zu berücksichtigen.

Benutzerorientierte Systeme unterstützen Benutzer und motivieren zum Lernen. Die Vorteile können erhöhte Produktivität, gesteigerte Arbeitsqualität, Verringerung der Neben- und Schulungskosten, sowie verbesserte Zufriedenstellung der Nutzer einschließen.

Die Entwicklung des Ansatzes soll unter Berücksichtigung von Aspekten der Kommunikationswissenschaften, Informatik und Mensch-Technik Interaktion geschehen. Dabei sollen die Vorteile existierender Methoden und Ansätze des klassischen Information Retrievals, der wissensbasierten Verfahren des Semantischen Netzes und der Informationsvisualisierung auf verschiedenen Ebenen erkannt, durch neuartige vervollständigt und zu dem in Abbildung 3.2 dargestellten Rahmenwerk integriert werden. Dieses Rahmenwerk soll durch eine synergistische Kombination von menschlichen und maschinellen Fähigkeiten Rechercheprozesse und die Synthese von gefundener und relevanter Information zu einem Gesamtbild und damit die Bildung von benötigtem Wissen in neuer Form unterstützen.

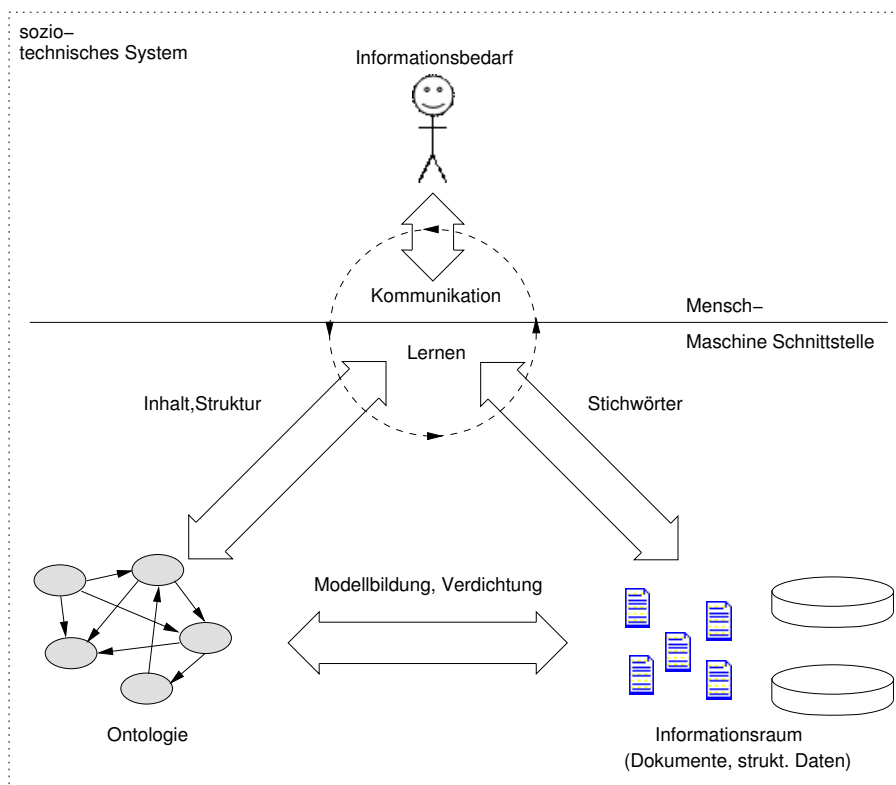


Abbildung 1.1: Rahmenwerk zur wissensbasierten Informationsrecherche

Es soll von der Annahme ausgegangen werden, dass in einer Recheresituation ein Kommunikationsprozess zwischen Mensch und Werkzeug stattfindet und dadurch ein soziotechnisches System durch die Kommunikationspartner gebildet wird. Erst diese Sichtweise ermöglicht die Annahme einer Vermittlung von Information, da diese nur innerhalb von Systemgrenzen direkt übertragbar ist. Die in diesem Kommunikationsprozess entstandenen Artefakte wie explizierte Präferenzen und persönliche Klassifikationssysteme, die sich in Lesezeichen zur Merkhilfe manifestieren, sollen zur Wissensbildung auf Seiten der maschinellen Verarbeitung genutzt und in einer Ontologie formalisiert werden. Diese systemische Wissensbasis soll dazu dienen, dem Nut-

zer individuelle Sichtweisen und verschiedenartige Recherchemöglichkeiten auf Basis semantischer Merkmale einzelner Inhalte zu ermöglichen. Darüber hinaus soll sie Hinweise auf bedeutungsvolle Zusammenhänge innerhalb des Informationsraums und zwischen einzelnen Inhalten liefern, sowie zu einer erleichterten Relevanzeinschätzung und Kontextualisierung von Ergebnissen beitragen. Damit soll auch das Verständnis des Informationsbedarfs der Nutzer gefördert werden, der das Ziel einer Recherche bestimmt.

Neben der Realisierung eines exemplarischen Prototypen zur Evaluation des Gesamtkonzepts soll ein weiterer Fokus der Arbeit auf der Gestaltung einer einheitlichen Nutzungsschnittstelle und Visualisierungstechnik liegen, welche die obige Kommunikation zwischen Mensch und Maschine überhaupt erst gestattet. Der Informationsraum und seine semantische Struktur soll dadurch sicht- und durchschaubar gemacht werden. Möglichkeiten sowohl zur unscharfen Stichwortsuche als auch zur präzisen semantischen Suche, sowie Funktionalitäten zum Stöbern auf Themen- und Ressourcenebene sollen integriert werden, um verschiedene und individuelle Sichten auf Information zu bieten. Dabei sollen Kontexte für die Explikation und Nutzbarmachung des Wissens einzelner Nutzer über bestimmte Inhalte und deren Zusammenhänge geschaffen werden.

Zusammengefasst sollen folgende, wesentliche Forschungsfragen beantwortet werden, die sich bei einer ontologiegestützten, semantischen Informationssuche ergeben:

- Auf welcher Ebene ist die ontologische Modellbildung eines Informationsraumes zur Informationssuche sinnvoll?
- Welche logische Architektur kann einer solchen Wissensbasis zugrunde gelegt und auf welche Art kann sie genutzt werden, damit verschiedenartige Rechercheprozesse unterstützt werden können?
- Wie kann eine Ontologie, neben herkömmlichen Methoden, aus den, in der Kommunikation mit dem System entstehenden Nutzerartefakten (Interaktionen, explizierte Präferenzen, Lesezeichen) generiert und erweitert werden?
- Wie kann sich eine Nutzerschnittstelle gestalten, welche die unterschiedlichen Arten der Informationssuche effektiv unterstützt und Einblicke in das eigene Informationsbedürfnis des Nutzers bietet?
- Wie kann die Nutzerschnittstelle die Konstruktion von ontologischen Abfragen für eine zielgerichtete Suche unterstützen?

Der zu entwickelnde Ansatz soll hierbei nicht als Konkurrenz zu bestehenden Verfahren verstanden werden, sondern vielmehr eine Erweiterung der Möglichkeiten zur allgemeinen und speziellen Informationsrecherche und persönlichen Wissensbildung darstellen.

1.3 Vorgehensweise

Die vorliegende Arbeit verfolgt die in Abbildung 1.2 gezeigten Vorgehensweise, um die formulierten Hauptziele der Arbeit zu erreichen.

Den Ausgangspunkt der Arbeit bildet Kapitel 2 mit der Untersuchung von bestehenden Methoden des Information Retrievals. Dazu werden zuerst in diesem Kontext häufig genutzte Begrifflichkeiten geklärt, sowie die Grundlagen und Vorannahmen der Arbeit geschildert. Darauf aufbauend wird der Methodenumfang beleuchtet, für den der Begriff Information Retrieval steht. Hierzu wird ein Überblick über die unterschiedlichen Ansätze geboten. Insbesondere werden die verschiedenen Methoden des merkmals- und wissensbasierten Retrievals gegenüber gestellt, sowie deren Vor- und Nachteile analysiert. Auch werden Visualisierungstechniken vorgestellt und diskutiert, die in diesem Kontext die Mensch-Computer Schnittstelle zu Information Retrieval Systemen bilden.

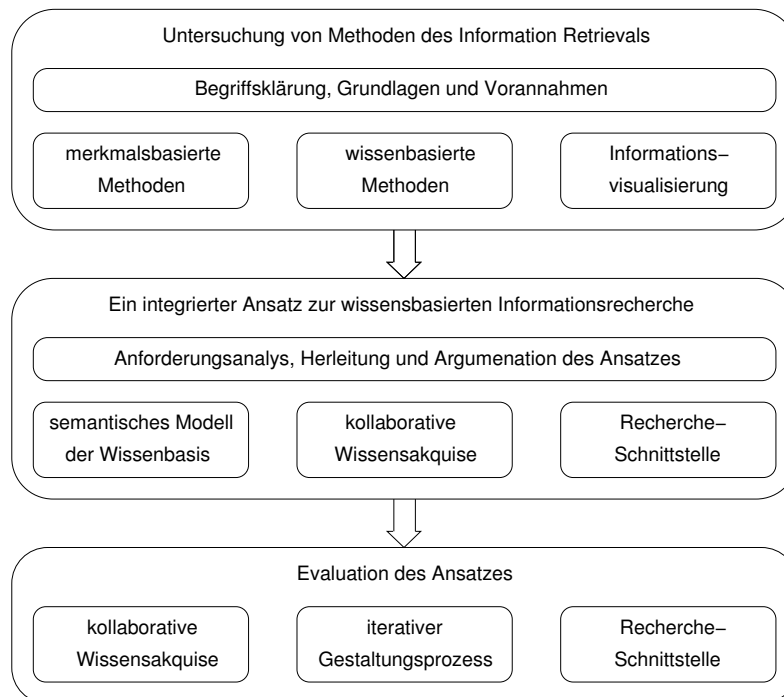


Abbildung 1.2: Aufbau der Arbeit

Auf Basis dieser Erkenntnisse folgt in Kapitel 3 zunächst eine Anforderungsanalyse, welche die Herleitung und Begründung des gewählten Ansatzes erlaubt. Es wird erkannt, welche existierenden Techniken in diesem Ansatz synthetisiert werden können und welche neu zu entwickeln sind. Diese werden in ein einheitliches Rahmenwerk integriert.

Im Einzelnen gilt es, eine Informationsarchitektur der ontologischen Wissensbasis zu entwerfen, die sowohl eine problemgerechte Modellierung eines textuellen Informationsraums als auch die Integration von nicht textuellen Inhalten erlaubt. Dabei kann es sich um Unternehmensdaten wie Personalprofile, Produktdaten und um gespeicherte Information von Datenbanksystemen handeln. Weiterhin wird ein Ansatz vorgestellt, der traditionelle Methoden der systemischen Wissensakquise mit einem kollaborativen Verfahren koppelt. Zur Unterstützung unterschiedlicher Suchstrategien bedarf es entsprechend angepasster Suchverfahren. Diese werden in einem Modell zusammengefasst. Das Kapitel schließt mit dem Entwurf der Schnittstelle zwischen Nutzer und System, mit der gewünschte Information im Dialog aufgefunden werden kann.

Eine Evaluation des Rahmenwerks erfolgt in Kapitel 4. Es erfolgt eine Bewertung der Qualität und Einsetzbarkeit der kollaborativen Wissensakquise aus Nutzerartefakten. Des Weiteren erfolgt die Darstellung der iterativen und nutzerzentrierten Entwicklung der Visualisierungs- und Interaktionswerkzeuge nach ISO 13407¹². Abschließend wird deren Integration in eine einheitliche Rechterschnittstelle innerhalb eines umfangreichen Nützlichkeits- und Benutzbarkeitstest bewertet.

Schließlich fasst Kapitel 5 die Ergebnisse der Arbeit zusammen und liefert einen Ausblick auf weitere Entwicklungsmöglichkeiten und Anwendungsszenarien. Diese Möglichkeiten beziehen sich auf die Optimierung der Nutzerschnittstelle, der technologischen Umsetzung zur besseren Skalierbarkeit und einer verbreiterten Einsatzmöglichkeit.

¹²INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: *ISO 13407: Human-centred Design Processes for Interactive Systems*, 1999

2 Methoden des Information Retrieval

"Verfahren der Erschließung und des gezielten Zugriffs auf gespeicherte Informationen"¹ bezeichnet in ihrer allgemeinsten Form die Anglizismen *Information Retrieval* (IR). Die Entwicklung und Optimierung solcher Verfahren ist, wie Eingangs erwähnt, keinesfalls neu, sondern das originäre Gebiet des Bibliothekswesens, welches auf eine immerhin 5000 Jahre alte Geschichte zurückblickt². Neu ist jedoch der seit der Digitalisierung der Bibliotheken und Unternehmen entstandene Versuch, diese zu automatisieren. Inzwischen werden dabei, aufgrund der Komplexität dieser Aufgabe, Erkenntnisse aus Forschungsbereichen der Ingenieurs- und Arbeitswissenschaften, der Informatik, Linguistik und sogar Philosophie mit einbezogen.

Ihre Anwendung finden diese Verfahren neben der erneuten Lokalisierung von bereits bekannter Inhalten in der Vermittlung von Information, das heißt im Zusammenhang des Lernens und der Wissensbildung. Ein IR-System ist im Vermittlungsprozess ein Werkzeug, mit dem bestimmte Inhalte selektiert werden können. Zur Interaktion mit diesem Werkzeug muss zum einen der Informationsbedarf des Nutzenden dem System in geeigneter Form übermittelt werden, zum anderen müssen die gefundenen Inhalte dem Nutzenden in geeigneter Darstellung präsentiert werden. Die Umsetzung des menschlichen Informationsbedarfs in eine für die Maschine verständliche Form und die Darstellung der maschinengerecht vorliegende Information auf eine menschlich wahrnehmbare Art sind neben der eigentlichen Suche zentrale Probleme des Information Retrieval³.

2.1 Begriffsklärung, Grundlagen und Vorannahmen

Einige, in diesem Kontext häufig verwendeter Begrifflichkeiten wie Daten und Information, sowie darauf aufbauend, Wissen, die Übertragung und Gewinnung von Wissen, Lernen und Kommunikation, erfreuen sich in der öffentlichen Diskussion als auch in einschlägiger Fachliteratur größerer Beliebtheit. Dies führt zu begrifflichen Unschärfen, so dass Missverständnisse und Verkürzungen nicht die Ausnahme, sondern fast die Regel sind. Vor der weiteren Präzisierung der Methoden des IR erfolgt daher zuerst eine Betrachtung und Abgrenzung dieser Termini.

2.1.1 Datum, Information und Wissen

Bei der Definitionen und Beschreibung der Zusammenhänge zwischen den Begriffen Daten, Information und Wissen wird gerne auf die Metapher einer Wissenspyramide⁴⁵ zurückgegriffen, deren unterste Ebene Daten und die Spitze Wissen darstellt:

"Die Zusammenhänge zwischen diesen Ebenen werden häufig als Anreicherungsprozess dargestellt. Zeichen werden durch Syntaxregeln zu Daten, welche in einem gewissen Kontext interpretierbar sind und damit für den Empfänger Information darstellen. Die Vernetzung von Information ermöglicht deren Nutzung in einem bestimmten Handlungsfeld, welches als Wissen bezeichnet werden kann. Teilweise werden aufbauend auf dieser Trennung noch zusätzliche Ebenen wie Weisheit, Intelligenz oder Reflexionsfähigkeit unterschieden."⁶

¹ LANGENSCHIEDT: *Fremdwörterbuch online*, 2004, Suchwort "Information Retrieval"

² vgl. BARTH: *5000 Jahre Bibliotheken - eine Geschichte ihrer Benutzer, Bestände und Architektur*, 1996

³ vgl. FERBER: *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*, 2003, S. 33

⁴ vgl. CHAFFEY: *E-Business and E-Commerce Management: Strategy, Implementation and Practice.*, 2001, S. 203

⁵ vgl. NORTH: *Wissensorientierte Unternehmensführung: Wertschöpfung durch Wissen*, 1998, S. 40

⁶ vgl. PROBST/RAUB/ROMHARDT: *Wissen managen: wie Unternehmen ihre wertvollste Ressource optimal nutzen*, 1997, S. 34-35

Die Bildung von Wissen erfolgt demnach über einen "Prozess der Bedeutungsanreicherung"⁷ von Daten und Information, der auch als Lernen verstanden werden kann. Daten sind dabei auf der syntaktischen Ebene anzusiedeln, während sich Information auf der semantischen Ebene befindet (vgl. Abbildung 2.1⁸). Dementsprechend enthalten also Datenbanksysteme nicht nur Daten, sondern auch Information, weil zusätzlich zu den Daten zumindest ein Teil der Semantik des jeweiligen Anwendungsgebietes auch im System modelliert ist⁹.

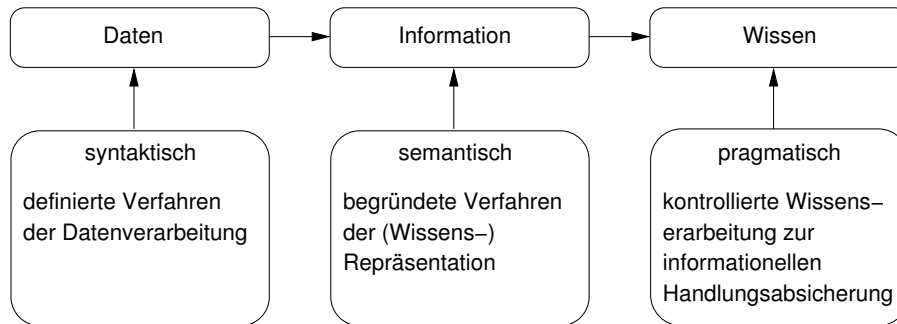


Abbildung 2.1: Daten - Information - Wissen

Wissen schließlich ist auf der pragmatischen Ebene definiert. In Abwandlung von Kuhlen¹⁰ lässt sich Wissen als die Teilmenge von Information definieren, die von jemandem in einer konkreten Situation zur Lösung von Problemen benötigt wird. Als Grundlage hierzu dienen begründete und wahre Überzeugungen¹¹.

Wie aus diesen Ausführungen hervorgeht, ist obige Definition von Wissen eng mit den Informationswissenschaften verknüpft. Bezogen auf technische Informationssysteme ist sie auch durchaus ausreichend, da es Maschinen prinzipiell nicht möglich ist, kognitive Prozesse zu vollziehen oder gar durch Reflexionsleistungen sinngemäß auf Daten einzuwirken. Da informationsverarbeitende Systeme jedoch im Sinne der Künstlichen Intelligenz (KI) der (wenn auch nur rudimentärer) Problemlösung mächtig sind, sei diese Definition beschränkt auf maschinelle Datenverarbeitung.

Bezogen auf den menschlichen Umgang mit Information wird eine einheitliche Definition des Begriffs Wissen erheblich schwieriger. Die Debatte darüber blickt auf eine zweitausendfünfhundertjährige Geschichte zurück, was die Schwierigkeiten, die damit verbunden sind, verdeutlicht. In der griechischen Philosophie spielt dabei gerade der Unterschied zwischen Wissen und Meinen eine entscheidende Rolle. Gegenüber dem bloßen Meinen unterscheidet sich Wissen durch die Angabe von Gründen in Bezug auf den zu erkennenden Gegenstand oder Sachverhalt in seinem Da- und Sosein. Wenn dieser Prozess der Angabe von Gründen wiederum einer Kritik unterzogen wird, spricht man von wissenschaftlicher Methode. Die Angabe von Gründen auf der Basis von Methodenwissen ist die Auszeichnung von Wissenschaft. Empirische Wissenschaft fragt nach dem Wie und Warum eines Sachverhaltes und stellt die Gründe in Form von gesetzmäßigen Zusammenhängen dar¹². Verifizierte Aussagen werden dann als Wissen anerkannt und vor (x-)beliebigen Aussagen ausgezeichnet¹³. Es kann sich allerdings erweisen, dass als wahr erkannte Aussagen sich im Fortlauf als unwahr erweisen. In diesem Falle wird von "unwahrem Wissen"¹⁴ gesprochen, was nicht, etwa wie bei Popper¹⁵, mit Nicht-Wissen gleichgesetzt werden

⁷SOUKUP: *Wissensmanagement: Wissen zwischen Steuerung und Selbstorganisation*, 2001, S. 223

⁸FUHR: *Information Retrieval*, 2004, S. 9

⁹vgl. FUHR: *Information Retrieval*, 2004, S. 8

¹⁰vgl. KUHLEN: *Zum Stand pragmatischer Forschung in der Informationswissenschaft*, 1990, S. 13-18

¹¹vgl. DELGRANDE/MYLOPOLOUS: *Knowledge Representation: Features of Knowledge*, 1986

¹²vgl. CAPURRO: *Wissensmanagement*, 2001, Abschnitt: "Was ist Wissen?"

¹³vgl. SCHREYÖGG/GEIGER: *Kann implizites Wissen Wissen sein?*, 2002

¹⁴vgl. LUHMANN: *Die Wissenschaft der Gesellschaft*, 1990, S. 273-293

¹⁵vgl. POPPER: *Die Logik der Sozialwissenschaften*, 1969

kann. Nicht-Wissen ist zunächst einmal die Beschreibung eines Wissensdefizits. Es handelt sich dabei jedoch nicht um eine fest beschreibbare Menge an fehlendem Wissen. Im Grundsatz ist Nicht-Wissen nur als mengentheoretische Negation von Wissen beschreibbar. Auf die Paradoxie, dass mit zunehmendem Wissen der wahrgenommene Umfang des Nicht-Wissens steigt, sei nur am Rande hingewiesen¹⁶¹⁷.

Der Wissensbegriff sei an dieser Stelle nicht weiter problematisiert. Es würde den Rahmen dieser Arbeit sprengen, sich dieser Debatte anzuschließen. Für das Verständnis von Information Retrieval und Systemen in diesem Anwendungsbereich ist es nicht unbedingt notwendig, alle Aspekte von Wissen zu erfassen. Es sei jedoch dargelegt, dass gemäß konstruktivistischer Weltanschauung, solches komplexes Wissen an psychische Systeme gebunden ist und nur anhand der Auseinandersetzung mit der Umwelt aufgebaut werden kann. Damit ist es nicht trivial übertragbar und lässt sich nicht auf bloße Speicherung oder Vernetzung von Information zu pragmatischer Handlungsfähigkeit reduzieren.

2.1.2 Vermittlung von Information durch Kommunikation

Wie beim Wissens- und Informationsbegriff existieren zur Vermittlung von Information und der Rolle der Kommunikation mehrere Ebenen der Betrachtung. So kommt im normalen Sprachgebrauch bei einem Zwiegespräch Kommunikation zustande. Dies impliziert zumindest einen Sender und einen Empfänger. Bei der Erklärung von Kommunikation wird daher meist auf die Informationstheorie von Shannon und Weaver¹⁸ zurück gegriffen. Diese befasst sich mit der Übertragung von Nachrichten von Sender zu Empfänger über (mehr oder weniger) gestörte Kanäle. Sie stellt ein mathematisches Modell für den Begriff Information und die Übertragung von Information in einem technischen Kommunikationssystem dar. Eine theoretische Vorhersage von Möglichkeiten und Grenzen der Informationsübertragung in Gegenwart von Störungen wird daher möglich.

Gerne wird dieses Modell auch auf menschliche Kommunikation übertragen, obwohl der semantische Aspekt einer Nachrichtenübertragung von Shannon und Weaver bewusst ausgeschlossen wurde. Im technischen Kontext der Shannon und Weaverschen Theorie ist die Bedeutung der Daten irrelevant. Jedoch könnte Bedeutung im Rahmen dieser Theorie sehr einfach berücksichtigt werden: Es wäre dazu nur ein weiterer (De-) Kodierungsschritt erforderlich, in dem einfach Encoder und Dekoder zweigeteilt anlegt sind. Bedeutung würde also zunächst in eine Form transformiert, die technisch übermittelbar wäre, dann übertragen und wieder dekodiert. Eventuelle Störungen wären dann nicht nur technischer, sondern auch semantischer Natur.

In den Sozialwissenschaften wird menschliche Kommunikation differenzierter betrachtet. Ausgangspunkt ist dabei nach Luhmann die Differenz von Mitteilung und Information. Information ist hierbei "ein Unterschied, der einen Unterschied macht"¹⁹. Dies bedeutet eine Änderung des internen Systemzustands der an der Kommunikation beteiligten Person. Wird eine Mitteilung zu Information, so wird sie von Luhmann als *anschlussfähig* bezeichnet. Mitteilungen sind anschlussfähig, wenn sie wahrgenommen und auf irgendeine Art und Weise weiter verarbeitet werden. Information existiert dabei nur als Ereignis im Bewusstsein, kann aber angenommen oder abgelehnt werden. Das Bewusstsein kann mit einem anderen Bewusstsein nur mit Hilfe von Mitteilungen kommunizieren, die sich als Verhalten manifestieren.

"Im Unterschied zu bloßer Wahrnehmung von informativen Ereignissen kommt Kommunikation nur dadurch zustande, dass Ego zwei Selektionen unterscheiden und diese Differenz seinerseits handhaben kann. [...] Die Differenz liegt zunächst in der Beobachtung des Alter durch Ego. Ego ist in der Lage, das Mitteilungsverhalten von dem zu unterscheiden, was es mitteilt. Wenn Alter sich seinerseits beobachtet weiß, kann er diese Differenz von Information und Mitteilungsverhalten selbst übernehmen

¹⁶LUHMANN: *Die Wissenschaft der Gesellschaft*, 1990

¹⁷POPPER: *Die Logik der Sozialwissenschaften*, 1969

¹⁸SHANNON/WEAVER: *The Mathematical Theory of Communication*, 1949

¹⁹LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S. 68

und sich zu eigen machen, sie ausbauen, ausnutzen und zur (mehr oder weniger erfolgreichen) Steuerung des Kommunikationsprozesses verwenden. [...] Das gibt dem Erwartungsbegriff für alle soziologischen Analysen eine zentrale Stellung."²⁰

Da durch die Verwendung von unterschiedlichen Kodes bei Sender und Empfänger die Mitteilung beim Empfänger nicht mehr zur gleichen Information werden kann, die sie beim Sender einmal war, entsteht nun Unsicherheit. Diese Unsicherheit, welche sich im Rahmen von Kommunikation zwischen zwei oder mehr Personen permanent ergibt, wird versucht, mit Hilfe der oben erläuterten Erwartung zu eliminieren²¹.

Das Konstrukt "Aufrichtigkeit" zeigt sehr gut, dass die Unzugänglichkeit des Kommunikationspartners zu Widersprüchen führt, wird gerade diese Unzugänglichkeit beziehungsweise die Differenz von Information und Mitteilung nicht eingeführt. Aufrichtigkeit wird gerade in dem Moment untergraben, in dem versucht wird, sie mitzuteilen: "Ich lüge nicht!?"²²

"Man kann gleichwohl nicht sagen, dass man meint, was man sagt. Man kann es zwar sprachlich ausführen, aber die Beteuerung erweckt Zweifel, wirkt also gegen die Absicht. Außerdem müsste man dabei voraussetzen, dass man auch sagen könnte, dass man nicht meint, was man sagt. Wenn man aber dies sagt, kann der Partner nicht wissen, was man meint, wenn man sagt, dass man nicht meint, was man sagt."²³

2.1.3 Soziotechnische Systeme

Ein System ist eine Abstraktion von Phänomenen und bezeichnet im Allgemeinen "eine abgegrenzte Anordnung von aufeinander einwirkenden Gebilden"²⁴. Solche Gebilde oder Elemente arbeiten in ihrer Verschiedenheit zusammen und erhalten als Ganzes ein bestimmtes Verhalten aufrecht. Durch die funktionalen Wechselwirkungen untereinander und die Abgrenzung zu einer Umwelt kann ein System als Einheit behandelt werden. Systeme können durch Differenzierung in einzelne Subsysteme aufgeteilt werden, wodurch das Gesamtsystem für diese wiederum als Umwelt fungiert. Anders herum wirken Subsysteme konstitutiv auf das Gesamtsystem. Existiert eine Wechselwirkung mit der Umwelt, so handelt es sich um offene Systeme, während autonome Systeme als geschlossen bezeichnet werden. In soziotechnischen Systemen bilden menschliche und technische Elemente eine Einheit.

Technische Systeme

Technische Systeme können als geschlossene Systeme angesehen werden und sind in erster Linie durch ihren Zweck charakterisiert. Zu deren definierenden Elementen zählen keine Menschen, wohl aber von Menschen geschaffene Artefakte. Diese Systeme lassen sich durch eine deterministische Relation zwischen Ursache und Wirkung beschreiben. Diese Relation, Übertragungsfunktion genannt, bestimmt die Identität eines technischen Systems im Gegensatz zu seiner Gegenständlichkeit. Auf eine gegebene Ursache erfolgt eine rekonstruierbare und reproduzierbare Abfolge von internen Zustandsänderungen, die eine bestimmte externe Wirkung oder Handlung erzielt. Handlungen dieser Art können wiederum durch andere technische Systeme ausgeführt werden. Der gesetzmäßige Zusammenhang zwischen Ursache und Wirkung ermöglicht eine externe und interne Steuerung und Überwachung.

²⁰ LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S. 198

²¹ Diesen Prozeß nennt Luhmann "doppelte Kontingenz".

²² PORR: *Eine interdisziplinäre Analyse von Niklas Luhmanns Werk*, 2002, S. 208

²³ LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S. 208

²⁴ DEUTSCHES INSTITUT FÜR NORMUNG: *DIN 19226: Regelungstechnik und Steuerungstechnik, Begriffe und Benennungen*, 1968

Soziale Systeme

Die Elemente *sozialer Systeme* bestehen nach Luhmann "aus Kommunikationen und aus deren Zurechnung als Handlung"²⁵. Zur Umwelt eines sozialen Systems gehören alle Elemente, die nicht im System selbst sind. Auch andere Systeme wie Maschinen, Organismen und psychische Systeme stellen für das betrachtete System die Umwelt dar. Die Einheit eines sozialen Systems existiert dabei nur aus der Sicht eines Beobachters. Der Beobachter stellt selbst ein System dar, welches aber Teil des beobachteten Systems sein kann. Durch diese Selbstbezüglichkeit ist ein soziales System seiner Systemgrenzen bewusst und kann diese auch beeinflussen²⁶. Die Systemumwelt kann nicht direkt beobachtet, jedoch durch deren Mitteilungen wahrgenommen werden. Daher besteht die Funktion eines sozialen Systems in der Konstruktion von Sinnzusammenhängen, die eine Selektion von systemexternen Mitteilungen erlauben. Dadurch wird eine Reduktion der äußeren Komplexität und Unsicherheit durch den Aufbau von Erwartungen erreicht. Auf eine gegebene Ursache erfolgt jedoch eine nicht deterministische Wirkung, womit sich ein solches System nicht steuern oder kontrollieren lässt. Innerhalb eines Autopoiesis genannten Prozesses, der "Produktion aus Produziertem"²⁷, erneuern sich soziale Systeme aus sich selbst. Konkret bedeutet dies, dass aus Kommunikationen weitere Anschlusskommunikationen entstehen. Die dabei übertragene Information kann angenommen oder abgelehnt werden. Luhmann unterscheidet das Interaktionssystem, die Organisation und die Gesellschaft als die drei primären Typen von sozialen Systemen. Ein Interaktionssystem wird als "Sozialsystem begriffen, das unter Anwesenden entsteht"²⁸ und stellt die kleinste Form eines sozialen Systems dar, welches sich nicht weiter als Subsystem ausdifferenziert. Darin werden Wahrnehmungsmöglichkeiten zur Verfügung gestellt und es ist nicht möglich, nicht darin zu kommunizieren. Eine Organisation ist ein soziales System, das als systembildende Operation Entscheidungen hat. Dabei sind Entscheidungen eine spezielle Form der Kommunikation, wenn aus zwei oder mehreren Handlungsalternativen eine ausgewählt wird. "Die Gesellschaft [wiederum, d.A.] ist ein kommunikativ geschlossenes System und kann nicht mit der Umwelt kommunizieren, sie findet dort niemanden, der ihr antworten könnte, und wenn, würde er eben dadurch in die Gesellschaft einbezogen werden"²⁹. Sie ist damit das umfassendste soziale System, das notwendigerweise keine Umwelt besitzt. Aufgrund der Selbstbezüglichkeit der Kommunikation sind soziale Systeme informatorisch und operational geschlossen. Daher kann von einer systeminternen Informationsübertragung gesprochen werden. Soziale Systeme benötigen jedoch mit Ausnahme der Gesellschaft eine Umwelt, welche die Ausbildung einer Systemgrenze erlaubt.

Synthese von technischen und sozialen Systemen

Im Unterschied zu rein sozialen Systemen beinhalten *soziotechnische Systeme* mindestens ein technisches Subsystem mit dem es in Interaktion steht. Der Begriff des soziotechnischen Systems geht auf einen Gestaltungsansatz zur gemeinsamen Optimierung sozialer und technischer Systeme des englischen Tavistock Institut zurück^{30,31}. Dadurch konnte die Notwendigkeit einer Integration von ingenieurmäßigem und verhaltenswissenschaftlichem Denken nachgewiesen werden. Der soziotechnische Gestaltungsansatz versucht die sozialen und technischen Anforderungen, Fragestellungen und Forderungen bei der Gestaltung und Entwicklung eines neuen Arbeitsgerätes aufeinander abzustimmen. Dabei wird vor allem auf die Integration sachlicher und menschlicher Aspekte im Rahmen der Zielfunktionen eines Unternehmens, der Ausdifferenzierung von Subsystemen mit eindeutig abgrenzbaren Aufgabenbereichen und einer Offenheit gegenüber der Systemumwelt Wert gelegt. Die Elemente eines soziotechnischen Systems sind Relationen zwischen Kommunikationsakten und technischen Kontrollhandlungen. Daraus ergibt

²⁵ LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S.240

²⁶ vgl. LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S. 618

²⁷ LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S. 233

²⁸ vgl. LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S. 535

²⁹ vgl. LUHMANN: *Soziale Systeme. Grundriss einer allgemeinen Theorie*, 1984, S. 549

³⁰ TRIST/BAMFORTH: *Human Relations 4 [1951]*, 1951

³¹ TRIST et al.: *Organizational Choice, Capabilities of Groups at the Coal Face under changing Technologies*, 1963

sich ein evolutionärer Prozess wechselseitiger Prägung. Kommunikationsakte seitens des sozialen Systems beziehen sich auf Kontrollhandlungen gegenüber dem technischen System. Diese Kontrollstrukturen prägen wiederum die Kommunikationsakte des sozialen Systems. Dabei ist die beiderseitige Kommunikation thematisch aufeinander bezogen und anschlussfähig.

Information Retrieval Systeme können damit als Interaktionssysteme betrachtet werden in denen sich die Subsysteme wechselseitig beobachten. In dieser Kommunikation lässt sich durch die informatorische Geschlossenheit Information zum sozialen System übertragen. Andersherum wird es aber auch dem technischen System ermöglicht, aus Beobachtung Information für sich zu generieren.

2.2 Einführung in das Gebiet des Information Retrieval

Gegenstandsbereich des *Information Retrieval* ist die Repräsentation, Speicherung, Organisation und Erschließung von Informationsräumen^{32,33}. Gewöhnlich wird allerdings unter Information Retrieval meist die Suche in oder nach Dokumenten verstanden³⁴, obwohl auch das Finden von (multi-)medialen Inhaltsträgern eingeschlossen ist. Multimediales Retrieval ist von den Betrachtungen in dieser Arbeit ausgeschlossen. Retrieval ist dabei nicht gleichbedeutend mit Recherche, sondern beinhaltet vielmehr die Gesamtheit der Methoden und Instrumente zur Informationsfindung, wohingegen unter Recherche lediglich der eigentliche Suchprozess zu verstehen ist³⁵. Ein Information Retrieval System umfasst vielmehr die methodischen Grundlagen, technische Verfahren und Einrichtungen, die das Retrieval ermöglichen³⁶.

Historisch gesehen ließe sich dieser Begriff also als eine Bezeichnung der Arbeiten von Cleverdon³⁷, Salton³⁸, Sparck-Jones³⁹ und Lancaster⁴⁰ sehen. Eine direkte Definition hierzu liefert Lancaster im Vorwort seines Buches:

"Information Retrieval ist die, wenn auch ungenaue Bezeichnung für die Arbeiten, die in diesem Buch besprochen werden. Ein Information Retrieval System unterrichtet nicht den Nutzer im Sinne einer Wissensveränderung über den Gegenstandsbereich seiner Anfrage. Es unterrichtet nur über die Existenz (oder Nichtexistenz) von mit der Anfrage in Verbindung stehenden Dokumenten mit und wo sie zu finden sind."⁴⁰

Dadurch werden Frage-Antwort Systeme^{41,42}, wie sie von Winograd und Minsk beschrieben werden, oder auch Expertensysteme^{43,44}, ausgeschlossen.

Inhaltlich abgegrenzt wird Information Retrieval auch gegen *Daten-Retrieval*, welches in der Regel nicht ausreichend ist, ein bestimmtes Informationsbedürfnis auszugleichen⁴⁵. Daten-Retrieval zielt auf das Auffinden aller Objekte ab, die wohl definierten Randbedingungen genügen. Dies ist zum Beispiel in regulären Ausdrücken oder in relationaler Algebra der Fall. Tabelle 2.1⁴⁶ zeigt die unterschiedlichen Merkmale der beiden Retrieval-Verfahren auf.

Die Fachgruppe Information Retrieval innerhalb der Gesellschaft für Informatik fasst diesen Begriff jedoch weiter und definiert ihre Ziele und Aufgaben wie folgt:

³²vgl. SALTON/MCGILL: *Introduction to Modern Information Retrieval*, 1983, Einleitung

³³vgl. BAEZA-YATES/RIBEIRO-NETO: *Modern Information Retrieval*, 1999, Einleitung

³⁴vgl. STAUD: *Wirtschaftsinformation*, 1997

³⁵vgl. POETZSCH: *Information Retrieval: Einführung in Grundlagen und Methoden*, 1998, S. 13

³⁶vgl. WORMSER-HACKER: *Evaluierung im Information Retrieval*, 1998, S. 13

³⁷vgl. CLEVERDON: *Progress in Documentation. Evaluation of Information Retrieval Systems*, 1970

³⁸vgl. SALTON: *Automatic Text Analysis*, 1970

³⁹vgl. SPARCK-JONES: *Automatic Keyword Classification for Information Retrieval*, 1971

⁴⁰übersetzt aus LANCASTER: *Information Retrieval Systems: Characteristics, Testing and Evaluation*, 1968

⁴¹WINOGRAD: *Understanding Natural Language*, 1972

⁴²MINSKY: *Semantic Information Processing*, 1968

⁴³FEIGENBAUM/BARR/COHEN: *The Handbook of Artificial Intelligence*, 1981

⁴⁴PUPPE: *Einführung in Expertensysteme*, 1991

⁴⁵vgl. BAEZA-YATES/RIBEIRO-NETO: *Modern Information Retrieval*, 1999

⁴⁶VAN RIJSBERGEN: *Information Retrieval*, 1979, S. 1

	Daten Retrieval	Information Retrieval
Ergebnisabgleich	exakt	partiell, bestmöglich
Inferenz	Deduktion	Induktion
Modell	deterministisch	probabilistisch
Klassifikation	monothetisch	polithetisch
Anfragesprache	formal	natürlich
Fragespezifikation	vollständig	unvollständig
gesuchte Objekte	die Fragespezifikation erfüllende	relevante
Reaktion auf Datenfehler	sensitiv	insensitiv

Tabelle 2.1: Daten- vs. Information Retrieval

"Im Information Retrieval (IR) werden Informationssysteme in Bezug auf ihre Rolle im Prozess des Wissenstransfers vom menschlichen Wissensproduzenten zum Informationsnachfragenden betrachtet. Schwerpunktmäßig werden jene Fragestellungen behandelt, die im Zusammenhang mit vagen Anfragen und unsicherem Wissen entstehen. Vage Anfragen sind dadurch gekennzeichnet, dass die Antwort a priori nicht eindeutig definiert ist. Hierzu zählen neben Fragen mit unscharfen Kriterien insbesondere auch solche, die nur im Dialog iterativ durch Reformulierung in Abhängigkeit von den bisherigen Systemantworten beantwortet werden können. Die Unsicherheit (oder die Unvollständigkeit) dieses Wissens resultiert meist aus der begrenzten Repräsentation von dessen Semantik (z.B. bei Texten oder multimedialen Dokumenten); darüber hinaus werden auch solche Anwendungen betrachtet, bei denen die gespeicherten Daten selbst unsicher oder unvollständig sind. Aus dieser Problematik ergibt sich die Notwendigkeit zur Bewertung der Qualität der Antworten eines Informationssystems, wobei in einem weiteren Sinne die Effektivität des Systems in Bezug auf die Unterstützung des Benutzers bei der Lösung seines Anwendungsproblems beurteilt werden sollte."⁴⁷

Als kennzeichnend für das Gebiet werden somit vage Anfragen mit unscharfen Kriterien und unsicherem Wissen betont. Die Art der Darstellung des Wissens ist dabei von untergeordneter Bedeutung, jedoch soll zur Bewertung und Auswahl der Nutzen für den Menschen herangezogen werden.

2.2.1 Qualitätsbeurteilung

Zur qualitativen Beurteilung von IR-Systemen wird zwischen *Effizienz* und *Effektivität* unterschieden. Unter Effizienz wird im Allgemeinen der möglichst sparsame Umgang mit Systemressourcen wie Speicherplatz, CPU-Zeit oder Antwortzeiten für die Bearbeitung einer bestimmte Aufgabe verstanden. Effektivität bezeichnet das Kosten-Nutzen-Verhältnis bei der Anwendung eines bestimmten Verfahrens. Bei der Nutzung eines IR-System bestehen die "Kosten" in dem vom Nutzer aufzubringenden Zeitaufwand und seiner mentalen Belastung bei der Lösung seines Problems mit Hilfe des Systems. Der erzielte Nutzen besteht in der "Qualität der erreichten Lösung"⁴⁸. Für die Beurteilung unter Effektivitätsgesichtspunkten wird meist der Begriff der *Relevanz* zugrunde gelegt.

Als Relevanzbeurteilung wird der Grad der Übereinstimmung zwischen einer Suchanfrage und der inhaltlichen Aussage eines Dokumentes aus der Treffermenge bezeichnet. Dabei geht die Spanne der intuitiven Relevanzbeurteilung eines Dokumentes durch den Nutzer von einer vollständigen Übereinstimmung mit dem gesuchten Thema über eine partielle, bis hin zu keinerlei Übereinstimmung mit dem Suchthema. Hinzu kommt noch die Betrachtung der gesamten Er-

⁴⁷FUHR: *Gesellschaft für Informatik: Ziele und Aufgaben der Fachgruppe „Information Retrieval“*, 1996

⁴⁸FUHR: *Information Retrieval*, 2004, S. 15

gebnismenge, in der sich mehrere, einzeln betrachtet nur partiell übereinstimmende, Dokumente gegenseitig so ergänzen, dass alle zusammen betrachtet den Informationsbedarf vollständig abdecken. Ebenso ist es möglich, dass mehrere, als relevant empfundene, Dokumente innerhalb einer Treffermenge im Wesentlichen den gleichen Inhalt aufweisen und somit teilweise redundant sind.

Als eines der charakteristischsten Merkmale zur Qualitätsbeurteilung des Information Retrieval gilt hierfür der klassische Ansatz zur Berechnung der Fähigkeit eines Retrieval-Systems, zu einem Suchthema aus der Dokumentenbasis alle passenden Dokumente nachzuweisen und im Vergleich dazu nur relevante und keine irrelevanten Dokumente zurückzuliefern. Diese beiden Kriterien werden mit *Vollständigkeit* (Recall) und *Genauigkeit* (Precision) bezeichnet⁴⁹.

Vollständigkeit und Genauigkeit: Ist $D = \{d_1, \dots, d_m\}$ eine Menge von Dokumenten, $q \in Q$ eine Anfrage und D_q die Menge der in D zur Anfrage q gefundenen Dokumente. Ist ferner R_q die Menge der zur Anfrage q relevanten Dokumenten, so berechnet sich die Vollständigkeit zu

$$R(q, D) := \frac{|D_q \cap R_q|}{|R_q|} \quad (2.1)$$

und die Genauigkeit zu

$$P(q, D) := \frac{|D_q \cap R_q|}{|D_q|} \quad (2.2)$$

Anschaulich betrachtet ist die Vollständigkeit der Quotient aus der Anzahl der gefundenen, relevanten Dokumente und der Anzahl aller relevanten Dokumente im Informationsraum. Ebenso ist die Genauigkeit das Verhältnis aus der Anzahl der gefundenen, relevanten Dokumente zu der Anzahl aller gefundenen Dokumente. Angemerkt sei an dieser Stelle, dass Relevanzbetrachtungen von IR-Systemen grundsätzlich sehr schwierig durchzuführen sind. Die Beziehung zwischen Informationsbedürfnis und Anfrage können sehr komplex sein und lassen sich nur schlecht auf eine Relevanzfunktion abbilden.

2.2.2 Komponenten eines Information Retrieval Systems

Prinzipiell kann eine IR Anwendung als ein System beschrieben werden, das aus einer Menge von Dokumenten und einer Menge von Suchanfragen besteht (vgl. Abbildung 2.2). Dabei sind die Dokumente und Suchanfragen in einer geeigneten Indexierungssprache repräsentiert. Ebenfalls enthält es einen Mechanismus, der die für eine Suchanfrage relevanten Dokumente bestimmt⁵⁰. Für die einzelnen Komponenten des Systems werden folgende Aufgaben gelöst:

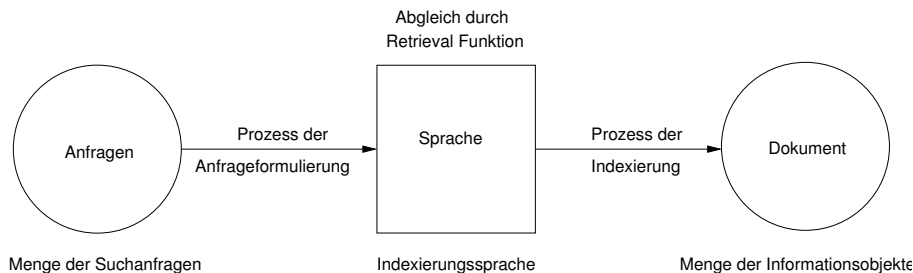


Abbildung 2.2: Grundschema des Information Retrievals

- **Informationserschließung:** Die Abbildung der gespeicherten Dokumente d auf deren Repräsentation d' übernimmt die Informationserschließungskomponente. Durch diesen, *Indexierung* genannten, Prozess werden Dokumente in eine Indexierungssprache übersetzt.

⁴⁹vgl. SALTON/MCGILL: *Information Retrieval - Grundlegendes für Informationswissenschaftler*, 1987, S. 172

⁵⁰vgl. SALTON/MCGILL: *Information Retrieval - Grundlegendes für Informationswissenschaftler*, 1987, S. 11

Dazu werden inhaltsbeschreibende Deskriptoren den Dokumente zugeordnet, so dass sie anhand dieser Merkmale gefunden werden können.

- *Abfragesprache*: Mittels einer Abfragesprache wird die Suchanfrage des Nutzers q in die entsprechende Repräsentation q' überführt. Dadurch kann durch die, aus der Frage gewonnene, Information über die Bedürfnisse des Nutzers, die dazu passenden Inhalte gefunden werden.
- *Informationsausgabe*: Die Informationsausgabe erfolgt durch eine Komponente zur Visualisierung der aufgrund der Anfrage gefundenen Suchergebnisse.

Den Abgleich der Anfragerepräsentation q' mit der Repräsentation d' einer Menge von Dokumenten nimmt dabei die normierte Retrieval-Funktion vor, die als eine Abbildung

$$R : q' \times d' \Rightarrow [0, 1] \quad (2.3)$$

angesehen werden kann.

Gerade der Indexierungsvorgang wird zu den wichtigsten, aber auch den schwierigsten Aufgaben im Rahmen des Information Retrieval gezählt⁵¹. Wichtigster Aspekt ist dabei, Begriffe oder Deskriptoren zu vergeben, die möglichst genau, den Inhalt eines Dokuments abbilden. Dabei wird zwischen *Stich-* und *Schlagwörtern* unterschieden. Stichwörter sind Terme, die dem Dokument zur Beschreibung entnommen werden. Schlagwörter sind Terme, die dem Dokument zur Inhaltsbeschreibung zugeordnet werden. Schlagwörter müssen nicht unbedingt im Dokument selbst enthalten sein. Erfolgt die Vergabe der Deskriptoren händisch, das heißt ohne maschinelle Hilfe, so wird von *manueller* beziehungsweise *intellektueller Indexierung* gesprochen. Im Falle der rein maschinellen Vergabe nennt sich der Vorgang *automatische Indexierung*. Auch Zwischenformen wie die *semiautomatische Indexierung* werden in der Literatur erwähnt⁵².

2.3 Manuelle Erschließungsverfahren

Manuelle Indexierungsverfahren stammen aus dem Bibliothekswesen, finden ihre Anwendung aber auch in wissensbasierten IR-Systemen (siehe Abschnitt 2.5). Zur Erschließung, sowie der tatsächlichen örtlichen Lokalisierung eines Mediums bedienen sich Bibliotheken so genannter *Kataloge*. Unter Katalog wird im allgemeinen Sprachgebrauch ein, nach bestimmten Gesichtspunkten (z.B. alphabetisch oder sachlich) geordnetes, Verzeichnis verstanden⁵³. An Katalogarten existieren der *Alphabetische Katalog*, der *Standortkatalog* und die beiden Sachkataloge, der *Schlagwortkatalog* und der *Systematische Katalog*.

Der Alphabetische Katalog verzeichnet Medien nach formalen Kriterien in alphabetischer Reihenfolge. Die formalen Elemente sind in der Regel der Verfassername, der Sachtitel und nach gegebenen Regeln der Name einer Körperschaft, die ein Werk erarbeitet oder an seinem Zustandekommen beteiligt war. Der Alphabetische Katalog beantwortet hauptsächlich die Frage, ob eine Bibliothek ein bestimmtes Buch besitzt, dessen Merkmale, wie Verfasser, bekannt sind.

Ein Standortkatalog führt Medien in genau der Reihenfolge auf, in der sie in den Bücherregalen aufgestellt sind - ein genaues Spiegelbild der Ordnung des Medienbestandes. Hiermit wird vor allem die Vergabe einer Signatur unterstützt. Jedem neuen Werk wird damit eine eindeutige Standortnummer zugeteilt, die den Ort des Mediums im Bestand festlegt. Außerdem bildet der Standortkatalog das Hilfsmittel für die von Zeit zu Zeit vorkommenden Revisionen des Buchbestandes.

Schlagwort- oder Stichwortkataloge erschließen den Bibliotheksbestand aufgrund des Inhalts eines Mediums. Diese, Sachkataloge genannten, Verzeichnisse sollen auf die Frage antworten, welche Werke eine Bibliothek über ein bestimmtes Thema oder Sachgebiet besitzt. Vom Aufbau her gleichen sie einem Lexikon und sind ebenfalls alphabetisch geordnet.

⁵¹vgl. SALTON/MCGILL: *Information Retrieval - Grundlegendes für Informationswissenschaftler*, 1987, S. 57

⁵²vgl. PANYR: *Information Retrieval Systeme: State of the Art*, 1987, S. 17

⁵³vgl. HACKER: *Bibliothekarisches Grundwissen*, 1992, S. 171

Im Gegensatz zu flachen Sach- und Stichwortlisten existiert als weiterer Sachkatalog der Systematische Katalog. Medien werden darin anhand ihres Inhalts in immer weiter verfeinerte Sachgebiete unterteilt. Er unterliegt ex verbis einer gewissen Systematik oder *Klassifikation*, ein weit verbreitetes Hilfsmittel, um in IR-Systemen Wissensgebiete zu organisieren. Der eigentliche Vorgang der Katalogisierung wird in der Regel von *Thesauren* durch die Bereitstellung eines kontrollierten Vokabulars zur Indexierung unterstützt.

2.3.1 Klassifikationen

Klassifikationen stellen ein "natürliches Ordnungsprinzip"⁵⁴ dar und erschließen Inhalte systematisch. Dabei werden meistens hierarchische Systeme verwendet, in deren Ebenen Themen unterschiedlich detailliert unterschieden werden. Strenge hierarchische Systeme können als Bäume dargestellt werden, wobei in der Wurzel alle Themen oder Objekte zusammengefasst werden. Die Blätter des Baumes werden durch einzelne Themen gebildet, welche nicht mehr weiter unterschieden werden. Die Bezeichnungen der detailliertesten Klassen kann als Pfad von der Wurzel bis zu einem Blatt dargestellt werden⁵⁵. Einer solchen Sortierung kann im Allgemeinen allerdings nur eine Sichtweise zugrunde liegen. So muss zum Beispiel entschieden werden, ob Autositze in die Klasse der Autoteile oder in die Klasse der Sitzmöbel fallen. Während die erste Einteilung den Aspekt, Teil von etwas zu sein, betont, wird bei der zweiten ein eher funktionaler Aspekt in den Vordergrund gerückt.

Nach Definition ist eine Klasse eine Gruppe von Dingen mit gemeinsamen Merkmalen. Formal lässt sich durch eine Klassifizierung, das heißt durch die Zugehörigkeit zu einer Klasse, "ein Attribut definieren, das genau diese Klasseneinteilung leistet"⁵⁶.

Klassifikation: Ist D eine Menge von Objekten. Eine Zerlegung von $D = K_1 \cup K_2 \cup \dots \cup K_d$ in paarweise disjunkte Teilmengen oder Klassen. $K_i \cap K_j = \emptyset, \forall i, j \in \{1, \dots, d\}, i \neq j$ heißt Klassifikation oder Klassifizierung der Objekte aus D .

Hierarchische Klassifikation: Bildet eine Menge H von Klassen K_1, K_2, \dots, K_d eine Halbordnung, so dass $\mathcal{H} = (H, >)$, so nennt sich \mathcal{H} eine hierarchische Klassifikation.

Strenge hierarchische Klassifikationssysteme werden gebildet indem deren Klassen weiter in disjunkte Teilmengen zerlegt werden. Neben strengen hierarchischen Klassifikationssystemen werden in der Praxis häufig auch Systematiken zugelassen, die nicht darauf bestehen, dass die Klassen disjunkt sind.

Eine Systematisierung dieses Ansatzes sind so genannte schwache Hierarchien oder Polyhierarchien, bei denen zugelassen ist, dass eine Klasse mehrere Oberklassen hat. Formal heißt das, dass bei der Bildung einer feineren Klassifikation nicht mehr verlangt wird, dass die neuen Klassen durch das Teilen von Klassen erzeugt werden. Ist ein Sachgebiet in solch einer Systematik als Untergebiet von mehreren Obergebieten eingetragen, wird auch von Doppelstellen gesprochen.

Klassifikationen in ihrer reinen Form zeichnen sich dadurch aus, dass sie a priori festgelegte Klassen enthalten. Wenn sie zur Einteilung von Medien genutzt werden, müssen die Klassen bereits definiert sein. Ein solches System wird als *präkoordiniert* bezeichnet. Es ist naturgemäß wenig flexibel und muss - um vollständig zu sein - von vorne herein sehr viele Klassen enthalten. Dadurch wird es unübersichtlich und erfordert, wie schon oben am Beispiel der Autositze gezeigt, oft schon auf sehr allgemeiner Ebene nicht rückgängig zu machende Entscheidungen, die nicht immer für alle Nutzenden nachvollziehbar sind.

Deshalb wurden zusätzliche Elemente entwickelt, mit denen Klassifikationen flexibler gemacht werden können. Das sind neben Doppelstellen so genannte Anhängeszahlen. Damit werden Modifikationen beschrieben, die sich häufig wiederholen und verhältnismäßig unabhängig von den Begriffen sind, auf die sie angewendet werden. Das können zum Beispiel Moderatoren sein, wie

⁵⁴ GAUS: *Dokumentations- und Ordnungslehre*, 1995

⁵⁵ vgl. FERBER: *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*, 2003, S. 47

⁵⁶ BROCKHAUS: *Der Brockhaus multimedial 2000 Premium*, 2000, Stichwort "Klasse"

die Konstruktion "Wartung von ..." oder "Handel mit ...". Diese Anhängeszahlen werden in der internationalen Dezimalklassifikation mit einem Strich an die Kennzahl des jeweiligen Begriffs angehängt. Neben diesen allgemeinen Begriffen existieren noch eine Reihe von Verknüpfungssymbolen, die ganz spezifische semantische Bedeutungen haben. Zum Beispiel wird durch das Gleichheitszeichen eine Sprache ausgedrückt: 860=20 bezeichnet "Spanische Literatur in englischer Sprache", durch runde Klammern ein Ort: 622.33(493) bezeichnet "Kohlebergbau in Belgien" oder durch die Anhängeszahl - 05 eine Person: 655.1 - 05 bezeichnet "Buchdrucker"⁵⁷. Durch solche Verfahren erhält die Klassifikation eine umfangreiche Syntax, mit der sich insbesondere komplexe Objekte genauer modellieren lassen. Sie gehen damit in Richtung eines Repräsentationsansatzes, wie er in der Künstlichen Intelligenz weiter entwickelt wurde. Diese Ansätze, Aspekte außerhalb der eigentlichen hierarchischen Klassifikation zu modellieren, indem Bezeichner erst bei der Einordnung eines Objekts konstruiert werden, werden als *Postkoordination* bezeichnet.

Neben der zeitinvarianten, thematischen Kategorisierung kann es sinnvoll sein, Inhalte anhand von Prozessen aufzuschlüsseln. Dies ist insbesondere dann geeignet, wenn in diesem Prozess spezifische Information an die Akteure für typische Aufgaben und Rollen vermittelt werden soll. Im funktionalen Kontext stellt eine Rolle das Aufgabenpaket einer Person dar⁵⁸, wobei sowohl mehrere Personen eine Rolle wahrnehmen können als auch eine Person verschiedene Rollen inne haben kann. Um die Rollen erstmalig zu erstellen, werden Prozesse formal analysiert und widerspruchsfrei abgebildet. Im nächsten Schritt werden ähnliche Aufgaben, oder solche, die gemeinsam ausgeführt werden müssen, zu Rollen zusammengefasst. Aufgaben können so als hierarchisch strukturierte Klassifikation betrachtet werden, auf deren oberster Ebene die Gesamtaufgabe steht. Diese kann auf mehreren Ebenen in weitere Elemente zerlegt werden. Dadurch wird sichergestellt, dass alle Aktivitäten in der richtigen Reihenfolge durchgeführt werden, aber auch, dass die Bearbeitung der zur Aufgabe benötigte Information rechtzeitig zur Verfügung steht⁵⁹. Dieser neuere Ansatz findet hauptsächlich bei dem Entwurf von Portalen zu betrieblichen Intranets Verwendung, kann aber auch auf Portale für Dienstleistungen der öffentlichen Hand übertragen werden. Dort wird gerne auf das, von den Sozialwissenschaften geprägte, Konzept der *Lebenslagen* zurückgegriffen. Es wird dazu benutzt, die Bürgerperspektive im Zusammenhang mit Behördengeschäften zu analysieren⁶⁰. Lebenslagen bezeichnen dabei spezielle Situationen im Leben einer Person oder Familie, welche einen gewissen Zeitraum überdauern und während denen Handlungen und Aktionen angestoßen werden. Anhand dieser Klassifikation könne zum Beispiel für die Lebenslage "Geburt" Inhalte zu allen Aktivitäten wie medizinische Untersuchungen, Ernährungsberatung, oder der Einkauf notwendiger Güter erschlossen werden.

Eine andere Weiterentwicklung der Klassifikationen hin zu mehr Flexibilität und einer stärkeren Ausdrucksfähigkeit sind *Facettenklassifikationen*. Hier werden zunächst Grundkategorien gebildet, die den Gesichtspunkten entsprechen, unter denen die Objekte betrachtet werden können. Für jede Grundkategorie werden dann, als mögliche Werte, so genannte Facetten angegeben. Diese Facetten können auch hierarchisch strukturiert sein⁶¹. Die Facettenklassifikation kann als mehrdimensionales System angesehen werden, bei dem in jeder Dimension eine, von den anderen Dimensionen (mehr oder weniger) unabhängige, Klassifikation angewendet wird. Dadurch wird eine größere Flexibilität erreicht. Facettenklassifikation gleicht dem Modell einer Faktendatenbank: In einer fest vorgegebenen Anzahl von Grundkategorien müssen Angaben gemacht werden, beziehungsweise es muss explizit angegeben werden, dass keine Angaben gemacht werden sollen.

Auf Seiten der Informationsrecherche ermöglichen diese Klassifikationen eine Suche zu erweitern oder einzuschränken. Wenn keine oder zu wenig Dokumente zu einem Begriff gefunden werden, kann durch die Auswahl allgemeinerer Oberbegriffe die Suche verbreitert werden. Umgekehrt kann die Menge der gefundenen Dokumente eingeschränkt werden, indem nur nach

⁵⁷Beispiele aus MANNECKE: *Klassifikation*, 1997

⁵⁸vgl. ESSWEIN: *Das Rollenmodell der Organisation: Die Berücksichtigung aufbauorganisatorischer Regelungen in Unternehmensmodellen*, 1993

⁵⁹vgl. KOPFERGER/SCHULTE: *Knowledge meets Process - Wissen und Prozesse managen im Intranet*, 2001, S.36

⁶⁰vgl. ALBER: *Versorgungsklassen im Wohlfahrtsstaat*, 1984

⁶¹vgl. MANNECKE: *Klassifikation*, 1997

bestimmten Unterbegriffen gesucht wird⁶².

Klassifikationen ermöglichen es durch ihre Systematik, Inhalte zu ordnen. Die Ordnungskriterien sind dabei aber häufig eher formal oder durch einen bestimmten Aspekt bestimmt. Damit sind sie in vielen Fällen wenig intuitiv und flexibel. Durch elektronische Systeme können komplexere, vielfältigere oder auch einfach mehrere parallele Systematiken angeboten werden, die im manuellen Betrieb nicht mehr zu bewältigen wären. Dadurch eröffnen sich weitere Zugriffswege zur Information.

2.3.2 Thesauren

In gewisser Weise bilden Thesauren das sprachliche oder terminologische Gegenstück zu hierarchischen Klassifikationssystemen. Ihr Schwerpunkt liegt allerdings mehr auf der Erfassung von Wörtern, Termen und Ausdrücken eines Sachgebiets und auf der Beschreibung der Beziehungen zwischen diesen, als auf der disjunkten Klassifikation von Objekten. Ihre Funktion besteht in der Definition eines kontrollierten Vokabulars und der Herstellung von Beziehung zwischen den Begriffen dieses Vokabulars. Dabei werden nicht nur, wie bei der Klassifikation, hierarchische Systeme aus Ober- und Unterbegriffen definiert, sondern es können zusätzlich eine Vielzahl von weiteren Beziehungen zwischen Wörtern dargestellt werden.

"Ein Thesaurus im Bereich der Information und Dokumentation ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederfinden dient. Er ist durch folgende Merkmale gekennzeichnet:

1. Begriffe und Bezeichnungen werden eindeutig aufeinander bezogen ("terminologische Kontrolle"), indem - Synonyme möglichst vollständig erfasst werden, - Homonyme und Polyseme besonders gekennzeichnet werden, - für jeden Begriff eine Bezeichnung (Vorzugsbenennung, Begriffsnummer oder Notation) festgelegt wird, die den Begriff eindeutig vertritt.
2. Beziehungen zwischen Begriffen (repräsentiert durch ihre Bezeichnungen) werden dargestellt."⁶³

Neben Homonymen und Polysemen verzeichnet ein allgemeiner Thesaurus, wie der von Roget⁶⁴, auch Ober- und mögliche Unterbegriffe auf. Thesauren dienen auf Autorensseite vor allem dazu, Kreativität und Vielfalt, aber auch Genauigkeit bei der Wortwahl zu unterstützen (oder vorzutäuschen). Im Bereich der Wiederfindung von Medien werden Angaben über deren Inhalt durch die Beschreibung mit definierten, vereinbarten und genormten Bezeichnungen erzeugt. In diesen Kernbestand von zugelassenen Ausdrücken für die Indexierung mit dem Thesaurus werden nur sorgfältig ausgewählte Terme aufgenommen, die eine wohldefinierte Bedeutung in dem zu beschreibenden Sachgebiet haben. Zu einem Deskriptor kann eine Menge von Synonymen angegeben werden, die in der Fachsprache zwar in der gleichen oder einer ähnlichen Bedeutung wie der Deskriptor verwendet werden, bei der Indexierung aber nicht zugelassen sind. Die Definition dieser Synonymmengen legt auch fest, wie genau ein Thesaurus die Details eines Sachgebiets aufschlüsselt. Soll wenig genau unterschieden werden, lassen sich verwandte Terme zu einer Synonymmenge zusammenfassen, soll dagegen eine sehr detailgenaue Beschreibung ermöglicht werden, dürfen nur Terme mit wirklich gleicher Bedeutung in einer Synonymmenge zusammengefasst werden. In vielen Fällen wird es sogar nötig sein, verschiedene Aspekte eines breit verwendeten Begriffs in verschiedene Thesaurusdeskriptoren aufzuspalten, um eine genügend hohe Spezifität des Thesaurus zu erreichen. In diesen Fällen müssen die spezifischen Bedeutungen durch weitere Definitionen oder Bemerkungen kenntlich gemacht werden.

Für Thesauren gibt es typischerweise zwei Zugangsarten: Zum einen gibt es eine alphabetische Liste mit den Deskriptoren. In diese Liste werden auch die Terme aus den Synonymmengen

⁶²vgl. HARMS/LUCKHARDT: *Virtuelles Handbuch Informationswissenschaft*, 2001, Exkurs "Information Retrieval"

⁶³DEUTSCHES INSTITUT FÜR NORMUNG: *DIN 1463-1: Erstellung und Weiterentwicklung von Thesauri*, 1987, S. 22

⁶⁴ROGET: *Thesaurus of English Words and Phrases*, 1852

aufgenommen. Von ihnen aus wird mit einer bestimmten Relation auf den zugelassenen Deskriptor der entsprechenden Synonymmenge verwiesen. Zum anderen existiert für Deskriptoren oder Ausdrücke, die aus mehreren Wörtern bestehen, einen so genannten rollierenden Index, in dem sie unter jedem der einzelnen Wörter, aus denen sie zusammengesetzt sind, alphabetisch zu verzeichnet sind.

Thesauren werden im Allgemeinen "von Hand" erstellt. Das kann zum Beispiel im Zusammenhang mit bibliographischen Arbeiten, wie Bibliographien, Bibliotheks- oder Museumskatalogen, Abstraktsammlungen und bibliographischen Informationsdiensten, aber auch zur Beschreibung von Wirtschaftsgütern und Produktionsbereichen im internationalen Handel geschehen. Dabei arbeiten meist mehrere Personen oder auch Institutionen zusammen, die mit dem Fachgebiet befasst sind, für das der Thesaurus konstruiert werden soll⁶⁵.

2.4 Automatische Erschließung und Retrieval Modelle

Die Methoden der automatischen Indexierung versuchen den Prozess der Informationserschließung, das heißt die Zuordnung von Deskriptoren zu den einzelnen Dokumenten, vollständig zu automatisieren, so dass die Funktion des menschlichen Indexierers entbehrlich wird⁶⁶. Auch wird, durch deren Integration in Retrieval-Modelle, der Prozess des Findens von Information unterstützt. Somit reagiert ein IR-System auf eine Anfrage mit einer Menge von Antwortdokumenten. In welcher Form jedoch Inhalte vorliegen (Texte, multimediale Dokumente, Bilder) und die damit einhergehende Wahl der Deskriptoren und Repräsentation (Vektoren, Records, Regeln, semantische Netze, ...) ist im Prinzip nicht beschränkt und so gibt es eine Fülle von Retrieval-Modellen. Sie unterscheiden sich hauptsächlich in Methoden, welche *exakte* oder *partielle Übereinstimmung* von Anfrage und Antwort erzielen. Letztere konzentrieren sich entweder auf bestimmte *Merkmale* oder auf die inhaltliche *Struktur* (vgl. Abbildung 2.3⁶⁷). Bezogen auf das Auffinden von

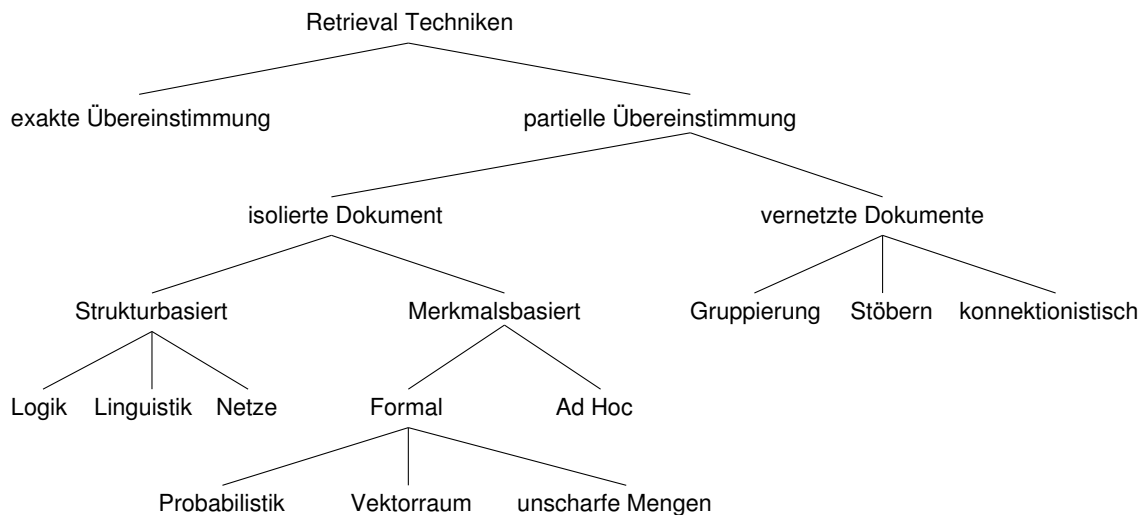


Abbildung 2.3: Klassifikation von Retrievalmodellen

Dokumenten seien diese im Folgenden skizziert⁶⁸.

⁶⁵vgl. FERBER: *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*, 2003, S. 33-35

⁶⁶vgl. KAISER: *Computer-unterstütztes Indexieren in Intelligenten Information Retrieval Systemen*, 1993, S.24

⁶⁷BELKIN/CROFT: *Retrieval Techniques*, 1987

⁶⁸vgl. BEKAVAC: *Information Retrieval*, 2001, S.16ff

2.4.1 Klassifikation von IR-Modellen

Die Methode der *exakten Übereinstimmung* qualifiziert alle Dokumente ohne Sortierung gleichwertig als Treffer wenn sie die Anfragebedingungen erfüllen. Bei nur geringfügigen Abweichungen werden sie als Treffer disqualifiziert, obwohl die systemgerechte Frageformulierung erkennbar nur eine grobe Näherung an den eigentlichen Informationsbedarf darstellt. Damit reduziert sich die Retrieval-Funktion auf die Abbildung $q' \times d' \rightarrow \{0 : 1\}$. Bei diesem, auch *Boolesches Retrieval*, genannten Verfahren formuliert der Nutzer mit Hilfe der booleschen Operatoren UND, ODER und NICHT iterativ seine Suchanfrage und lässt sich dabei im Wesentlichen von den Trefferanzahlen, sowie den inspizierten Antwortdokumenten leiten. Zu viele Treffer erzwingen eine Einschränkung der Anfrage, zu wenige Treffer erfordern eine Erweiterung der Fragestellung.

Bei den Methoden mit *partieller Übereinstimmung* wird die Antwortmenge einer Anfrage gemäß der entsprechenden Retrieval-Funktion R (2.3) nach vermuteter Relevanz sortiert (Ranking). Eine initiale Anfrage kann daher so spezifisch wie möglich gestellt werden. Statt der leeren Menge - wenn keine von ihnen die spezifischen Anforderungen erfüllt - werden die am besten passenden Dokumente als Ergebnis angeboten. Weiter wird bei partieller Übereinstimmung unterschieden, ob einzelne Dokumente oder vernetzte Dokumente Gegenstand des Retrieval-Prozesses sind. Im Falle der Betrachtung einzelner Dokumente wird zwischen strukturbasierten und merkmalsbasierten Ansätzen unterschieden. Strukturbasierte Ansätze verwenden meist Methoden der Künstlichen Intelligenz (KI) und werden in Abschnitt 2.5 eingehender behandelt.

Merkmalbasierte Ansätze versuchen Dokumente anhand bestimmter Gewichtungen von spezifischen Eigenschaften wie Begriffen, Phrasen oder sonstigen Kennzeichen zu erfassen. Je nach deren Repräsentation (Wortvektor, Netz) finden verschiedene Vergleichsverfahren Anwendung. Die nach Belkin&Croft als formal bezeichnete Ansätze repräsentieren solche Mengen probabilistisch, vektoriell oder als unscharfe Menge.

Probabilistische Modelle versuchen, durch Betrachtung von Häufigkeiten gewisser Merkmale auf die Wahrscheinlichkeit zu schließen, dass ein Dokument d auf eine Anfrage f als relevant einzustufen ist. Wurde eine Menge von Dokumenten vom Benutzer als relevant eingestuft, so kann eine Wahrscheinlichkeit abgeleitet werden, dass andere Dokumente ebenfalls relevant sind, wenn diese die Merkmale aus den schon als relevant beurteilten Dokumenten beinhalten.

Im Vektorraum-Modell werden die Merkmale der Dokumente und Anfragen als Vektoren repräsentiert, die Anzahl aller Dokumente spannt einen n -dimensionalen Vektorraum auf. Durch den Suchanfragevektor werden als Ergebnismenge diejenigen Dokumente ermittelt, deren Dokumentvektoren sich in räumlicher Nähe zu dem Suchanfragenvektor befinden. Die räumliche Nähe wird dabei mit Ähnlichkeitsmaßen, wie etwa dem eingeschlossenen Winkel der Vektoren, dem Kosinus-Maß, bestimmt.

Die Repräsentation von Dokumenten durch die Logik mit unscharfen Mengen nach Zadeh⁶⁹ gilt als eine Erweiterung der zweiwertigen booleschen Algebra auf Mehrwertigkeit. Bei jedem Merkmal wird dabei über eine Art Gewichtung, die durch eine Funktion definiert wird, die Zugehörigkeit zu einer Merkmalsklasse bestimmt. Die Ergebnismenge wird dann über die Auswertung der Merkmalsfunktionen der Dokumente und der Suchanfrage bestimmt.

Der als ad hoc bezeichnete merkmalsbasierte Ansatz verwendet ebenfalls numerische Verfahren zum Vergleich von Mengen. Dabei werden allerdings keine eindeutigen Modelle verwendet, sondern Verfahren kombiniert, die zu dem besten Ergebnis führen. Die so "ad hoc" entstandenen Techniken sind dadurch aber nicht immer exakt nachvollziehbar.

Sind vernetzte Informationsräume Gegenstand des Retrievals, so kommen netzwerkbasierete Verfahren zum Einsatz. Der älteste Ansatz hierzu ist das Gruppierungsverfahren. Als Gruppe wird hierbei eine Menge von Dokumenten ähnlichen Inhalts definiert. So werden alle verfügbaren Dokumente verschiedenen Gruppen zugeordnet, wobei ein Dokument mehreren Gruppen angehören kann. Wird ein Dokument für eine Suchanfrage als relevant empfunden, so gehören die Dokumente derselben Gruppe ebenfalls zu der Ergebnismenge. A priori vorhandene Vernetzung ist also keine Voraussetzung für das Gruppierungsverfahren, vielmehr werden Dokumente durch Gruppierung erst untereinander in Beziehung gesetzt. Die partielle Übereinstimmung von Doku-

⁶⁹ZADEH: *Fuzzy Sets and Systems*, 1965

ment und Anfrage bezieht sich hier eher auf die Gewichtung von zugehörigen Anhäufungen als auf die Relevanzabschätzung einzelner Dokumente.

Stöbern bezeichnet den navigatorischen Suchvorgang in Systemen, in denen Dokumente oder Teile davon netzwerkartig durch Verknüpfungen verbunden sind. Vor allem das Retrieval in Hypertextsystemen wird mit diesem Ansatz in Verbindung gebracht.

Ähnlich wie beim Stöbern bilden untereinander verknüpfte Knoten die Grundlage für den konnektionistischen Ansatz⁷⁰⁷¹. Die Knoten beinhalten hierbei jedoch die inhaltliche Erschließung von Dokumenten in Form von Begriffen. Die Vernetzung erfolgt von Begriffen zu anderen, ähnlichen Termini und zu den Dokumenten selbst. Eine Suchanfrage startet bei einem relevanten Knoten und prüft rekursiv von diesem aus, entlang seiner Verknüpfungen, andere Knoten. Auf ähnliche Weise arbeiten auch bayessche Inferenznetze⁷².

Strukturbasierte Ansätze des Information Retrievals verwenden inhaltliche Elemente zur Repräsentation der Dokumente. Das beinhaltet linguistische Strukturmerkmale oder die logisch fundierte Modellierung der inhaltlichen Struktur mittels semantischer Netze, Ontologien oder Frames. Diese Ansätze werden auch wissensbasierten Verfahren genannt (siehe Abschnitt 2.5). Linguistische Ansätze versuchen, die in den Dokumenten enthaltenen Terme nicht als bloße Zeichenketten aufzufassen, sondern als bestimmte Formen eines Wortes. Mittels Lemmatisierung durch Thesauren können die verschiedenen Flexionsformen eines Wortes, deren Synonyme, Antonyme, Super- und Subbegriffe bei der Suche mitberücksichtigt werden.

Im tatsächlichen Anwendungsfall werden die vorgestellten Ansätze meist untereinander kombiniert. Anders herum lassen sich die in IR- Systemen verwendeten Techniken meist nicht einem bestimmten Ansatz eindeutig zuweisen. Vor allem beim Umgang mit vernetzten Informationsräumen und bei Ansätzen, bei denen eine kleinere relevante Dokumentmenge Voraussetzung für das weitere Vorgehen ist (z.B. Gruppierungs- und probabilistische Verfahren), ist eine Kombination verschiedener Retrieval-Techniken üblich.

Nach diesem kurz gefassten Überblick sollen im Weiteren in der Praxis noch immer häufig verwendete Retrievaltechniken genauer vorgestellt werden. Trotz ihres Alters bilden sie die Grundlage moderne IR-Systeme. Dabei finden nachstehende Abkürzungen Verwendung:

$D = \{d_1, d_2, \dots, d_d\}$	bezeichnet die Menge aller Dokumente
$T = \{t_1, t_2, \dots, t_m\}$	stellt das zur Indexierung verwendete Vokabular dar
$Q = \{q_1, q_2, \dots, q_k\}$	benennt die Menge aller Anfragen
\vec{d}_d, \vec{q}_d	ist das in eine geeignete Darstellung überführte Dokument/Anfrage
d'_d, q'_d	ist das als Vektor von Indextermen repräsentierte Dokument/Anfrage

2.4.2 Termgewichtung

Merkmalbasierte Methoden, welche partielle Übereinstimmung von Anfrage und Systemantwort ermöglichen, machen einen extensiven Gebrauch der Statistik. Zur Einschätzung der inhaltlichen Beschreibungsfähigkeit bestimmter Wörter in Bezug auf ein Dokument wird unterstellt, dass die Bedeutung dieser Wörter eng mit der Häufigkeit ihres Auftretens zusammenhängt. Die Gewichtung der Häufigkeit dieser Wörter soll zur Indizierung der Dokumente führen. Diese Gewichtung und die Auswahl der Wörter sind dabei so zu erheben, dass einerseits thematisch ähnliche Dokumente erkannt, als auch das Unterscheiden von solchen mit unterschiedlichen Themengebieten ermöglicht wird. Diese Forderungen an das Indexvokabular werden auch *Spezifität* und *Exhaustivität* genannt⁷³.

Als einer der Pioniere der automatischen Indexierung gilt H.P. Luhn⁷⁴. Bei ihm werden die absoluten Häufigkeiten der Wörter in einer Dokumentsammlung berechnet und nur solche zur Indexierung zugelassen, welche öfters als ein unterer und seltener als ein oberer Schwellwert

⁷⁰ COHEN/KJELSDEN: *Information Retrieval by Constraint Spreading Activation in Semantic Networks*, 1987

⁷¹ CRESTANI: *Application of spreading activation techniques in Information Retrieval*, 1997

⁷² TURTLE/CROFT: *Inference Networks for Document Retrieval*, 1990

⁷³ vgl. VAN RIJSBERGEN: *Information Retrieval*, 1979, S. 14

⁷⁴ LUHN: *A statistical Approach to mechanized Encoding and Searching of Library Automation*, 1957

auftauchen. Das Ergebnis dieser Vorschrift lässt sich als lokaler Gewichtungsfaktor einer normierten Termfrequenz $w_l(t_m) = \text{tf}(t_m) = \frac{h(t_m)}{h(t_{max})}$ auffassen. Dabei wird die Häufigkeit des Terms t_m durch die Häufigkeit des Terms t_{max} , welcher am meisten in einem Dokument auftritt, dividiert.

Andere Gewichtungen werden aus der Verteilung eines Indexterms in der ganzen Dokumentenkollektion abgeleitet. Gewichtsverfahren dieser Art werden als globale Gewichtungseinflüsse in der inversen Dokumentenfrequenz $w_g(t_m) = \text{idf}(t_m) = \ln\left(\frac{|D|}{d(t_m)}\right) + 1$ zusammengefasst, wobei $d(t_m)$ die Anzahl der Dokumente, in denen der Term t_m auftritt und $|D|$ die Anzahl aller Dokumente bezeichnet⁷⁵. Der Gewichtungswert w_t fällt mit wachsendem $d(t_m)$ monoton. Der Logarithmus dämpft große Werte und schwächt in diesen Formeln die Gewichte seltener Terme wieder ab.

Zusammenfassend lässt sich sagen, dass lokale Gewichtungseinflüsse eher die Spezifität und der Einfluss globaler Gewichtung eher die Exhaustivität des Indexvokabulars bestimmt⁷⁶. Es liegt daher nahe, beide Gewichtungen zu vereinen. Daraus entsteht das weit verbreitete tf-idf genannte Maß,

$$w(t_m) = \text{tf-idf}(t_m) = \frac{h(t_m)}{h(t_{max})} \cdot \ln\left(\frac{|D|}{d(t_m)}\right) + 1 \quad (2.4)$$

welches ursprünglich von Salton and Yang⁷⁷ entwickelt wurde. Sie kamen dabei zu einigen Erkenntnissen. Begriffe mit hoher, totaler Termfrequenz eignen sich unabhängig von ihrer Verteilung in den Dokumenten nicht sehr gut zur Indexierung. Terme mit mittlerer Auftretenshäufigkeit eignen sich am besten, besonders wenn ihre Verteilung in den Dokumenten verzerrt ist. Seltene Wörter mit verzerrter Verteilung sind auch geeignet, aber nicht so gut wie solche mit mittlerer Häufigkeit. Ganz seltene Begriffe, sowie diejenigen, mit hoher, totaler Termfrequenz sind ebenfalls gut geeignet, sollten aber gegenüber den vorher genannten benachteiligt werden. Da aber der experimentelle Beweis dieser Untersuchungen nicht ausreichend ist, können keine genaueren Aussagen gemacht werden⁷⁸.

2.4.3 Boolesches Retrieval

Boolesches Retrieval ist historisch als erstes Retrievalmodell entwickelt und eingesetzt worden und stellt noch oft den Standard bei vielen kommerziellen IR-Systemen dar, insbesondere bei digitalen Bibliotheken⁷⁹. Die Grundidee des Booleschen Retrieval ist es, Mengenoperationen auf Mengen von Objekten anzuwenden, die durch Indexterme charakterisiert sind. Es beschränkt den Suchvorgang im Wesentlichen auf das Faktenretrieval, das heißt darauf, zu prüfen, ob eine wohldefinierte Bedingung, wie das Auftreten einer bestimmten Zeichenkette in einem Text, erfüllt ist oder nicht. Eine Anfrage hat in diesem Modell die Form eines aussagelogischen Terms, wobei die Indexterme typischerweise mit den booleschen Operationen UND, ODER und NICHT verknüpft werden. Meistens können die einzelnen Begriffe links oder rechts trunkiert werden. Linkstrunkierung ermöglicht die Berücksichtigung von Komposita (Haus-Musik -> Hausmusik), während Rechtstrunkierung die Auffindung von Flexionsformen (Hausmusik->Hausmusiker) ermöglicht. Es können durch Sonderzeichen alle Terme bezeichnet werden, die ein bestimmtes Zeichenmuster in einer bestimmten Position enthalten. Indexterme sind hierbei ungewichtete Stichwörter. Die Repräsentation eines Dokuments aus einer Kollektion D mit $k = |T|$ Indexterme aus einem Indexvokabular T besteht aus einem binären Vektor $d'_d = \vec{d}_d$ mit $d_{d_i} = \{0, 1\}$ für $i = 1, \dots, k$. Jedes Wort $q_m = q'_m$ aus der Menge des Indexvokabulars T kann zur Anfrage q verwendet werden. Die Menge aller Anfragen Q lässt sich also aus $q_i \in T$ mit $q_1 \wedge q_2$, $q_1 \vee q_2$ oder $\neg q_i$ konstruieren. Die Retrieval-Funktion R wird rekursiv entsprechend der angegebenen Regel zu

⁷⁵ SPARCK-JONES: *A statistical Interpretation of Term Specificity and its Application in Retrieval*, 1972

⁷⁶ vgl. VAN RIJSBERGEN: *Information Retrieval*, 1979, S. 14

⁷⁷ SALTON/YANG: *On the Specification of Term Values in automatic Indexing*, 1973

⁷⁸ vgl. VAN RIJSBERGEN: *Information Retrieval*, 1979, S. 15

⁷⁹ z.B. OPAC - Online Public Access Catalog, Bibliothekskatalog im Online-Zugriff mit Recherche-, Bestell-, Verlängerungs- und Vormerkfunktionen.

$$R_B(q_i, \vec{d}_m) := \begin{cases} 1 & : q_i \in d_m \\ 0 & : \neg q_i \in d_m \\ \min(R(q_1, \vec{d}_m), R(q_2, \vec{d}_m)) & : q_1 \wedge q_2 \\ \max(R(q_1, \vec{d}_m), R(q_2, \vec{d}_m)) & : q_1 \vee q_2 \\ 1 - R(q_i, \vec{d}_m) & : \neg q_i \end{cases} \quad (2.5)$$

gebildet. $R_B(q_i, \vec{d}_m)$ hat den Wert 1, wenn q_i in dem Dokument d_m enthalten ist und ansonsten den Wert null. Ein theoretischer Vorteil der booleschen Anfragesprache besteht in ihrer Mächtigkeit. Es kann gezeigt werden, dass mit einer booleschen Anfrage jede beliebige Teilmenge von Dokumenten aus einer Datenbasis selektiert werden kann. Voraussetzung ist dabei, dass alle Dokumente unterschiedliche Indexierungen besitzen. Dieser theoretische Vorteil ist aber von geringer praktischer Bedeutung, da ein Nutzer in der Regel nicht genau weiß, wie die zu seiner Frage relevanten Dokumente aussehen. Daher kann er auch die Anfrage nicht entsprechend der hier skizzierten Vorgehensweise formulieren.

Um einige der Nachteile des Booleschen Retrievals zu überwinden, wurde eine Erweiterung vorgeschlagen, die auf der Theorie der Logik mit unscharfen Mengen⁸⁰ basiert. Im Unterschied zum booleschen Modell werden hier bei den Dokumentrepräsentationen nun auch gewichtete Indexierungen zugelassen. Die Beschreibung der Anfrage q und der Retrieval-Funktion R_B sind wie beim Booleschen Retrieval definiert. Durch die gewichtete Indexierung liefert die Retrieval-Funktion allerdings jetzt Werte $R_B(q_i, \vec{d}_m) \in [0, 1]$. Damit ergibt sich im Gegensatz zum booleschen Modell nun eine Rangordnung der Antwortdokumente und die diesbezüglichen Nachteile des Booleschen Retrieval entfallen. Theoretische Überlegungen wie auch experimentelle Untersuchungen zeigen aber, dass die Definition der Retrieval-Funktion ungünstig ist. Der Grund hierfür ist die Verwendung der Minimums-Funktion \min bei der konjunktiven Verknüpfung zweier Suchbegriffe. Auch wenn einige der Suchbegriffe eine hohe Relevanz in Bezug auf Antwortdokumente signalisieren, wird immer nur die niedrigste Gewichtung zurückgeliefert⁸¹.

2.4.4 Vektorraummodell

Das Vektorraummodell (VRM) ist wahrscheinlich das bekannteste Modell aus der IR-Forschung. Es wurde ursprünglich im Rahmen der Arbeiten am SMART-Projekt entwickelt⁸². SMART ist ein experimentelles Retrievalsystem, das von Gerard Salton und seinen Mitarbeitern seit 1961 entwickelt wurde. Dabei handelt es sich nicht nur um ein einzelnes System, sondern um eine Experimentierumgebung, in der eine Vielzahl von Verfahren und Algorithmen getestet wurden.

Im VRM werden die Repräsentationen der Dokumente und Anfragen als Punkte in einem Vektorraum aufgefasst, der durch die Indexterme des Informationsraums aufgespannt wird. Beim Retrieval wird dann nach solchen Dokumenten gesucht, deren Vektoren im Sinne einer vorgegebenen Metrik ähnlich zum Fragevektor sind. Durch diese geometrische Interpretation ergibt sich ein sehr anschauliches Modell. Der zugrunde liegende Vektorraum wird als orthonormal angenommen. Alle Term-Vektoren sind orthogonal, damit auch linear unabhängig und normiert. Die im VRM zugrunde gelegte Repräsentation eines Dokuments ist eine gewichtete Indexierung.

Dieser Index enthält für jedes Dokument d_d die gleichen Indexterme t_m und kennt zu jedem Paar (d_d, t_m) eine normierte Gewichtung w_l , mit $0 \leq w_l \leq 1$. Daraus ergibt sich für jedes Dokument ein m -dimensionaler Indexvektor \vec{d}_d mit $m = |T|$. Er hat exakt so viele Dimensionen, wie es Indexterme gibt. In jeder Zeile des Vektors \vec{d}_d steht als Eintrag die Gewichtung w_m des jeweiligen Indexterms t_m für das Dokument d_d . Im Vektorraum lassen sich diese Vektoren als Zeilen der so genannten *Term-Dokument-Matrix*

$$W = \{w_{i,j}\}_{i=1\dots m; j=1\dots n} \quad (2.6)$$

⁸⁰ZADEH: *Fuzzy Sets and Systems*, 1965

⁸¹vgl. FUHR: *Information Retrieval*, 2004, S. 59ff

⁸²SALTON: *The Smart Retrieval System - Experiments in Automatic Document Processing*, 1971

auffassen. Diese Matrix oder Relation enthält alle für das IR-System verfügbare Information einer Dokumentsammlung, die sich mit der Repräsentation durch Terme beziehungsweise Merkmale darstellen lässt.

Systemanfragen werden ebenfalls als Vektoren aufgefasst. Der Anfragevektor \vec{q}'_k ist genau so aufgebaut, wie der Indexvektor \vec{d}'_n eines Dokumentes. In jeder Zeile von \vec{q}'_k steht eine Gewichtung, die ausdrückt, in wie weit der entsprechende Indexterm in den Dokumenten des Anfrageergebnisses berücksichtigt werden soll.

Zur Berechnung der Antwortmenge auf eine Anfrage bestimmt nun ein IR-System für jedes Dokument dessen Relevanz anhand der Ähnlichkeit der Anfrage- und Dokumentvektoren. Im einfachsten Falle wird hierbei das Skalarprodukt $\vec{d}'_n \times \vec{q}'_n$ aus den beiden Vektoren berechnet. Beim Skalarprodukt hängen die Ähnlichkeitswerte jedoch stark von Anzahl und Größe der einzelnen Werte im Vektor die ungleich 0 sind, das bedeutet von der Anzahl der Terme in der Anfrage oder im Dokument ab. Längere Dokumente haben statistisch gesehen daher größere Chancen, hohe Ähnlichkeitswerte zu bekommen, als kurze. Daher schlagen Jones und Furnas⁸³ nach dem Vergleich mehrerer Ähnlichkeitsmaße das Kosinus-Maß als Ähnlichkeitsfunktion vor. Die Ähnlichkeitswerte sind damit unabhängig von der euklidischen Länge der Vektoren, also der Anzahl und Größe der von 0 verschiedenen Einträge und nur bestimmt von ihrer Richtung. Die Retrieval-Funktion ergibt sich damit zu

$$R_{V_{RM}} := \text{sim}_{\cos}(\vec{d}'_n, \vec{q}'_n) = \frac{\vec{d}'_n \times \vec{q}'_n}{|\vec{d}'_n| |\vec{q}'_n|} \quad (2.7)$$

2.4.5 Probabilistische Modelle

Die probabilistischen Verfahren betrachten die Relevanz einer Systemantwort auf eine Anfrage als Zufallsgröße. Es wird versucht, auf die Wahrscheinlichkeit zu schließen, inwieweit ein Dokument d auf eine Anfrage q als bedeutend einzustufen ist. Sind diese Wahrscheinlichkeiten für jede Anfrage und alle Inhalte gegeben, so können sie als Sortierkriterium der Elemente der Systemantwort genutzt werden. Da diese Wahrscheinlichkeit aber nicht unmittelbar zugänglich ist, wird versucht sie zu schätzen. Das probabilistische Retrieval Modell basiert daher auf der Voraussetzung, dass eine Menge von Dokumenten zuvor schon auf gegebene Anfragen als bedeutsam eingestuft wurde. Daraus werden mit entsprechenden Vereinfachungen und Unabhängigkeitsannahmen die a priori Wahrscheinlichkeiten abgeleitet. Terme, die bei einer gegebenen Anfrage zuvor schon in relevant eingestuft Dokumenten enthalten waren, erhalten eine höhere Gewichtung als Terme, die nicht in diesen relevanten Dokumenten vorkamen.

Im Modell von Robertson und Spark-Jones⁸⁴ wird zunächst ein Ereignis R eingeführt, welches besagt, dass ein bestimmtes Dokument d als relevant zu einer Anfrage q einzustufen ist. Ebenfalls wird eine m -dimensionale Zufallsvariable $\vec{X} = \{x_1, x_2, \dots, x_m\}$ mit $m = |T|$ eingeführt. Die Elemente x_m können die Werte 0 oder 1 annehmen, je nachdem, ob der Anfrageterm q'_m in \vec{d}'_m enthalten ist ($q'_m \in \vec{d}'_m$) oder nicht. Damit ergibt sich die Retrieval-Funktion R als bedingte Wahrscheinlichkeit zu

$$R_P := P(R|d, q) = P(R|\vec{X}) \quad (2.8)$$

welche sich unter Zuhilfenahme des bayesschen Theorems

$$P(X|Y) = \frac{p(X \cap Y)}{p(Y)} \quad (2.9)$$

berechnen lässt. Allerdings nur unter der Annahme der Unabhängigkeit zwischen Terme und Relevanz, womit sich Gleichung 2.9 zu $P(X|Y) = \frac{p(X) \cdot p(Y)}{p(Y)} = p(X)$ vereinfacht. Das probabi-

⁸³JONES/FURNAS: *Pictures of Relevance: A Geometric Analysis of Similarity Measures*, 1987

⁸⁴vgl. ROBERTSON/SPARK-JONES: *Relevance Weighting of Search Terms*, 1976

listische Modell lässt sich auch zur Gewichtung von Indextermen heranziehen. Robertson und Spark-Jones entwickeln dafür die Gleichung

$$w_P(t_m) = \ln \frac{\frac{R(q,m)+0.5}{R(q)-R(q,m)+0.5}}{\frac{d(t_m)-R(q,m)+0.5}{|D|-d(t_m)-R(q)+(Rq,m)+0.5}} \quad (2.10)$$

wobei $|D|$ die Anzahl der Dokumente in der Sammlung bezeichnet, $R(q)$ die Anzahl der zur Anfrage q relevanten Dokumente in einer Trainingsmenge, $d(t_m)$ die Anzahl der Dokumente, die den Term t_m enthalten und $R(q, m)$ die Anzahl der relevanten Dokumente, die den Term t_m enthalten. Im Zähler steht das Verhältnis der Anzahl der relevanten Dokumente, die den Term t_m enthalten zur Anzahl der relevanten Dokumente, die den Term nicht enthalten. Die Addition von 0.5 verbessert die Schätzung und sorgt dafür, dass nicht durch 0 dividiert wird. Im Nenner steht das Verhältnis der Anzahl der Dokumente, die den Term enthalten und nicht relevant sind zur Anzahl derer, die den Term nicht enthalten und nicht relevant sind⁸⁵.

Die Ähnlichkeit des probabilistischen Modells zu statistischen Verfahren wird deutlich, wenn keine a priori Relevanzeinschätzungen seitens Nutzer vorhanden sind. Croft und Harper⁸⁶ machen dazu die Annahme, alle Anfrageterme hätten die gleiche Wahrscheinlichkeit in relevanten Dokumenten vorzukommen. Formal wird dabei $R(q)$ und $R(q, m)$ zu 0 gesetzt. Mit Hilfe einer Grenzwertbetrachtung ergibt sich dadurch genau die invertierte Dokumentfrequenz $\text{idf}(t_m) = \ln \frac{|D|-d(t_m)}{d(t_m)} = \ln \frac{|D|}{d(t_m)} + 1$.

Im probabilistischen Modell lässt sich leicht eine Relevanzrückkopplungsschleife (Relevance Feedback) durch Einbeziehung des Nutzers erreichen⁸⁷. Nach einer Anfrage kann der Nutzer eine Bewertung des Systemvorschlags vornehmen. Anhand dieser Bewertung wird vom IR-System eine erneute Gewichtung der Antwortmenge vorgenommen. Je nach System können dadurch auch Termgewichtungen dauerhaft verändert werden.

2.4.6 Diskussion

Der Prozess der Informationsvermittlung ist, wie Eingangs gezeigt, grundsätzlich von Ungenauigkeit und Unsicherheit geprägt. Wenn im Kontext des Information Retrievals von unsicherer Information oder Ungenauigkeit gesprochen wird, so ist in der Regel nicht wirklich Information im nachrichtentechnischen Sinn, sondern deren Interpretation gemeint. Auf Seiten der Informationssuchenden sind die Anfragen dadurch gekennzeichnet, dass die Antwort im Vorfeld nicht eindeutig definiert werden kann. Unsicherheiten des Informationsträgers resultieren aus ungenauem und unvollständigem Wissen über den Bedarf des Informationssuchenden und der Informationsbasis. Erst durch einen Dialog von Informationsträger und -konsument und dessen Interpretation lassen auf das tatsächliche Informationsbedürfnis und die zu vermittelnde Information schließen.

Traditionell ist diese Arbeit das Aufgabengebiet von Bibliotheken. Der Bibliothekar in seiner Rolle als Mediator erbringt diese Interpretationsleistung und liefert mit Hilfe seiner Ausbildung und den Instrumentarien zur inhaltlichen Erschließung wie Systematiken und Thesauren die gewünschte Information. Jedoch besteht die Hauptproblematik der intellektuellen Erschließung in dem hohen Zeitaufwand, den hohen fachlichen Anforderungen an die Qualifikation des Indexierers und den damit verbundenen hohen Kosten. Bezogen auf die Verfügbarkeit von Information wird der Abstand zwischen dem Zeitpunkt der Veröffentlichung und der Suchbarkeit eines Dokuments infolge der großen Informationsmenge und deren exponentiellen Wachstumsrate, besonders in Intranets und Wissensdatenbanken, immer größer. Dies berührt im betrieblichen Kontext vor allem das informationslogistische Paradigma: Richtige Information muss aktuell und zur richtigen Zeit am richtigen Ort vorliegen. Aus der schier unendlichen Informationsmenge ergibt sich auch ein qualitatives Problem. Die Einheitlichkeit der Vergabe der Deskriptoren kann nicht gewährleistet werden, da die Person des Indexierers mit großer Wahrscheinlichkeit immer wieder wechselt.

⁸⁵FERBER: *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*, 2003, S. 123

⁸⁶CROFT/HARPER: *Using Probabilistic Models of Document Retrieval without Relevance Information*, 1979

⁸⁷vgl. ROCCHIO: *Document Retrieval Systems - Optimization and Evaluation*, 1966

Auch bei ein und derselben Person ist nicht sichergestellt, dass bei Dokumenten gleichen Themas dieselben Deskriptoren vergeben werden. Organisatorische Schwierigkeiten entstehen bei Eingliederung von Indexierern in ein Unternehmen. Einerseits ist das Indexieren zwar keine Managementfunktion sondern eine informationstechnische Unterstützungstätigkeit, andererseits ist auch die Zuordnung zur Datenverarbeitungsabteilung problematisch, da die Tätigkeit des Indexierens auch keine typische Datenverarbeitungstätigkeit ist. Durch mangelnde organisatorische Kontrolle wächst die Wahrscheinlichkeit der inkonsistenten Vergabe von Deskriptoren⁸⁸.

Dem gegenüber stehen die vielzähligen Methoden der automatisierten Erschließung und Vermittlung. Grosse Informationsbestände können mit diesen in relativ kurzer Zeit und zu niedrigen wirtschaftlichen Kosten orts- und zeitunabhängig suchbar gemacht werden. Das Hauptproblem des automatischen Information Retrievals besteht jedoch in der Interpretation und Abbildung von menschlichen Informationsartefakten in ihre maschinenverstehbare Repräsentation, bei der, wie bei jeder Modellbildung und Abstraktion, ein prinzipieller Informationsverlust immanent ist.

Merkmalsbasierte Ansätze stützen sich auf das gleichzeitige Auftreten von Termen in einer Anfrage und in einem Dokument. Die inhaltliche Bedeutung wird durch Indexterme modelliert, deren jeweiliges Gewicht die Relevanz die Gesamtbedeutung des Textes angibt. Dieses Gewicht wird vor allem durch die Analyse der Worthäufigkeit gewonnen. Ein Begriff, der in einem Dokument oft vorkommt, muss deswegen aber nicht ein guter und inhaltsrelevanter Deskriptor sein. Diese starke Orientierung an der Worthäufigkeit kann unter anderem auch dazu führen, dass die einzelnen Dokumente durch ihre Deskriptoren nicht gut voneinander unterschieden werden⁸⁹. Weitere Mängel bei der inhaltlichen Erschließung ergeben sich aus der Vernachlässigung des Wortkontextes und der Wortform, sowie aus Problemen bei der Erkennung von Mehrwortbegriffen, Synonymen und Polysemen. Im probabilistischen Modell wird zusätzlich die Annahme der statistischen Unabhängigkeit von Termen innerhalb der Sprache gemacht. Diese Annahme scheint im Allgemeinen recht unrealistisch.

Durch Einbeziehung einer Nutzerbewertung als Relevanzrückkopplung zur Neugewichtung der Indexterme kann jedoch eine deutliche Qualitätssteigerung verzeichnet werden⁹⁰. Ein besonders eindrucksvolles Verfahren zur Sortierung von Suchergebnislisten nach Relevanz liefert die Internet-Suchmaschine Google mit ihrem PageRank⁹¹ genannten Sortierungsverfahren. Webseiten werden auf ihre Vernetzungsstruktur untersucht und abhängig von der Anzahl der auf eine jeweilige Seite zeigenden Verweise ein Relevanzmaß abgeleitet. PageRank kann jedoch nur bei vernetzten Dokumenten Verwendung finden.

Ein weiterer Unsicherheitsfaktor ergibt sich aus der Anfrageformulierung. Oft korrespondieren die Suchterme nicht mit den Begriffen in einem relevanten Dokument, oder sie werden in gefundenen Dokumenten in einer unterschiedlichen Bedeutung verwendet⁹². Vom Nutzenden wird daher eine gewisse Kenntnis des Suchraumes und des darin verwendeten Vokabulars gefordert oder aber er benötigt die Hilfe Dritter.

2.5 Wissensbasiertes Information Retrieval

Strukturbasierte Ansätze zum Information Retrieval konzentrieren sich, im Gegensatz zur Betrachtung von Inhalten als bloße Menge von Merkmalen, auf inhaltliche Elemente zur Repräsentation und Wiederauffindung. Grundlage hierzu ist das Vorhandensein von systemischem Hintergrundwissen, welches Interpretation und Schlussfolgerung zulässt. Historisch ließe sich sagen, dass das Gebiet des wissensbasierten Information Retrievals aus der Überlappung der Forschung in der Künstliche Intelligenz (KI) und der Forschung auf dem Gebiet des Information Retrieval heraus entstanden ist⁹³.

⁸⁸vgl. KAISER: *Computer-unterstütztes Indexieren in Intelligenten Information Retrieval Systemen*, 1993, S. 21

⁸⁹vgl. KAISER: *Computer-unterstütztes Indexieren in Intelligenten Information Retrieval Systemen*, 1993, S. 49

⁹⁰vgl. SPARCK-JONES: *Reflections on TREC*, 1995

⁹¹PAGE et al.: *The PageRank Citation Ranking: Bringing Order to the Web*, 1998

⁹²vgl. BELKIN: *Anomalous States of Knowledge as a Basis for Information Retrieval*, 1980

⁹³vgl. CROFT: *Approaches to Intelligent Information Retrieval*, 1987

Aus einer anwendungsorientierten Sicht lassen sich viele Aufgaben der KI in die Gebiete natürlichsprachliche Systeme, Deduktionssysteme, bildverarbeitende Systeme, Robotertechnologie und Expertensysteme einordnen. Die meisten dieser Aufgaben verlangen die Anwendung von Wissen im Sinne einer Erfassung, Speicherung und Verarbeitung von relevanter Probleminformation. Daher werden Systeme, die in diesen Bereichen entwickelt werden, wissensbasierte Systeme genannt⁹⁴. Dieses Verständnis von wissensbasierten Systemen steht im Gegensatz zu deren Betrachtung als "... Werkzeuge und Mittel, die den Mitarbeitern [einer Organisation; d.A.] zur Verfügung gestellt werden, um Informationen abzulegen, aufzufinden und Wissen zu kommunizieren"⁹⁵.

Im Information Retrieval besitzen hauptsächlich die Bereiche natürlichsprachliche Systeme aber auch, nicht erst seit der Idee des Semantischen Netzes, Deduktionssysteme Relevanz. Dabei soll zur der Beschreibung eines Dokuments nicht nur die syntaktische Struktur, sondern auch Bedeutung erfasst werden. Durch Interpretation der Anfrage und dem semantischen Verständnis der Informationsbasis, kann so ein IR-System die Vollständigkeit und Relevanz seiner Antwortmenge erhöhen. Sparck-Jones definiert ein wissensbasiertes IR-System folgendermaßen:

"Ein wissensbasiertes Retrieval System ist ein System mit einer Wissensbasis und der Möglichkeit, Schlussfolgerungen zu bilden, welche dazu benutzt werden können, Verbindungen zwischen einer Anfrage und einer Menge von Dokumenten herzustellen."⁹⁶

Die Inkorporation von Wissen kann gemäß Abbildung 2.3, S. 34 auf zwei verschiedene Arten erfolgen. Linguistische Ansätze nutzen morphologisches und semantisches Wissen über Wörter, um algorithmisch den regelhaften Charakter der gesprochenen Sprache abzubilden. Logikbasierte Methoden modellieren Sachverhalte, unabhängig von einer bestimmten sprachlichen Ausdrucksform, auf Grundlage von Aussagen. Die nach Belkin und Croft als Netze benannten Verfahren stellen eher Formalismen zur Modellierung maschinellen Wissens dar.

2.5.1 Ontologien und semantische Netze

Grundlage zur Anwendung eines jeden wissensbasierten Systems ist die vorausgehende begriffliche Erfassung der entsprechenden Anwendungsdomäne und dessen Formalisierung in eine ausführbare Wissensbasis. Im Sinne der KI sind auch die Möglichkeiten der gemeinsamen Nutzung und Wiederverwendung solchen Domänenwissens anzustreben. Im Bereich der formalen Wissensrepräsentation wurden daher Ontologien als zentrale Gestaltungsobjekte mit hohem Wiederverwendungspotential vorgeschlagen^{97,98,99}.

Der Begriff "Ontologie" wurde aus der Philosophie entlehnt und bezeichnet dort die Lehre vom Seienden, seinen Eigenschaften im Allgemeinen, seinen Arten und Stufen, dem Verhältnis zum Sein, zum Dasein und zum Wesen. Seit Aristoteles wird sie als erste Philosophie, als Prinzipienwissenschaft angesehen. Ursprünglich als Metaphysik bezeichnet, bekam die Ontologie im 17. Jahrhundert, vor allem durch Christian Wolff, diesen Namen. In heutiger Zeit ist damit zugleich das Bekenntnis zu einer Philosophie verbunden, die eine sichere Seinsgrundlage wissenschaftlicher und philosophischer Aussagen zugrunde legt, die demzufolge die prinzipielle Bedeutung der Erkenntnislehre Kants bestreitet.

Im Kontext der KI wird eine Ontologie mit der am häufigsten zitierte Definition von Gruber definiert. "Eine Ontologie ist eine explizite Spezifikation einer Begriffsbildung"¹⁰⁰. Interessant ist hierbei die Vergegenständlichung einer Ontologie zu einem Objekt, welches Begriffsklassen, deren Hierarchien, Individuen (Instanzen), typisierten Relationen und Attribute zur Möglichkeit der

⁹⁴ vgl. BÜNING/LETTMANN: *Logik und Regelverarbeitung*, 2002, S. 2f

⁹⁵ BULLINGER/MÜLLER/RIBAS: *Wissensbasierte Informationssysteme - Enabler für Wissensmanagement*, 1999, S. 21

⁹⁶ übersetzt aus CROFT: *Approaches to Intelligent Information Retrieval*, 1987, S. 249

⁹⁷ NECHES et al.: *Enabling Technology for Knowledge Sharing*, 1991

⁹⁸ FARQUHAR/FIKES/RICE: *The Ontolingua Server: A Tool for collaborative Ontology Construction*, 1996

⁹⁹ USCHOLD/GRÜNINGER: *Ontologies: Principles, Methods, and Applications*, 1996

¹⁰⁰ übersetzt aus: GRUBER: *A Translation Approach to Portable Ontology Specifications*, 1993, S. 1; "Begriffsbildung" im Englischen "Conceptualization" von "Concept" im Sinne von Begriff

Schlussfolgerung formal abbildet¹⁰¹. Studer et al. fassen den Begriff Ontologie daher etwas enger zu "... einer formalen und expliziten Spezifikation einer gemeinsamen Begriffsbildung"¹⁰². Ontologien stellen damit einen formalen, also maschinenverstehbaren Repräsentations- und Austauschformalismus begrifflichen Wissens dar, der es den beteiligten Subjekten ermöglicht, egal, ob Mensch oder Maschine, ein gemeinsames Verständnis einer Diskurswelt zu bilden und auf Basis dieser miteinander zu kommunizieren (vgl. Abschnitt 2.1.2, S. 24). Entscheidendes Merkmal dabei ist die Definition einer einheitlichen Semantik, was implizit durch die inhaltliche Definition verschiedener struktureller Merkmale geschehen kann, wie zum Beispiel die Bildung von Klassenkonstruktoren, oder durch die explizite Angabe von Axiomen, das heißt immer wahren Aussagen. Am Rande sei bemerkt, dass hier eine Diskurswelt immer nur einen kleinen Ausschnitt der realen Welt beinhaltet, die zur Verwendung in einem wissensbasierten System modelliert wird.

Im Hinblick auf Wiederverwendung und gemeinsamer maschineller Nutzung solcherart explizierten und formalisierten Wissens können Ontologien auch im Sinne von softwaretechnischer Trennung in einem Klassenmodell zwischen der Ausführungsschicht, das heißt der Implementierung und ihrer Schnittstelle betrachtet werden¹⁰³.

Ontologien können auf unterschiedlichen Abstraktionsstufen modellbildend sein. Fensel unterscheidet dabei folgende Ebenen¹⁰⁴:

- Domänenspezifische Ontologien repräsentieren Terminologien, die innerhalb eines bestimmten Bereichs wie beispielsweise Elektronik, Medizin oder Mechanik gültig sind.
- Generische Ontologien versuchen den gesunden Menschenverstand in Bezug auf die Welt als solches abzubilden. Sie stellen also allgemeine Begriffe und Konzepte wie Zeit, Raum oder Ereignisse zur Verfügung¹⁰⁵. Als Konsequenz davon sind sie in mehreren Sparten gültig.
- Metadaten-Ontologien, wie zum Beispiel Dublin Core¹⁰⁶, stellen ein Vokabular zur Auszeichnung von Inhalten bereit.
- Repräsentationsontologien beziehen sich nicht auf eine spezifische Domäne, sondern liefern begriffliche Einheiten zur Definition von Repräsentationssprachen. Eine bekannte Ontologie dieses Typs ist die *Frame Ontology*¹⁰⁷, welche Konstrukte wie Frames, Slots und Slot Constraints definiert. Hiermit lässt sich terminologisches Wissen objektorientiert oder framebasiert darstellen (siehe auch Frame-Logik in Abschnitt 2.5.1.1).
- Methoden- und Aufgabenontologien^{108/109} definieren ein domänenunabhängiges Vokabular als Basis zur Beschreibung von Problemlösungskonzepten sowie Aufgabentypen.

Im Kontext von Ontologien fallen durch ihre strukturelle Darstellbarkeit als Graph sowohl in der wissenschaftlichen als auch in der populären Literatur gerne die Begriffe *semantische Netze* und *Themenkarten*. Semantische Netze wurden ursprünglich von Quillian als Repräsentation zur reichhaltigeren Ausgestaltung der Beziehungen der Terme in Thesauren vorgeschlagen¹¹⁰. Ein bekannter Vertreter semantischer Netze ist WordNet¹¹¹, ein seit 1985 in Princeton entwickeltes System von Synonymmengen, die durch typisierte Relationen verknüpft sind. Themenkarten sind nach ISO/IEC Standard 13250:2000¹¹² standardisiert und bieten die Möglichkeit der Formulierung von semantischen Netzen. Im *World Wide Web* (WWW) werden Themenkarten durch den

¹⁰¹ vgl. GUARINO: *Formal Ontology and Information Systems*, 1998, S. 3

¹⁰² übersetzt aus: STUDER/BENJAMINS/FENSEL: *Knowledge Engineering: Principles and Methods*, 1998, S. 162

¹⁰³ vgl. GRUBER: *A Translation Approach to Portable Ontology Specifications*, 1993, S. 2

¹⁰⁴ FENSEL: *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*, 2000

¹⁰⁵ NOY/HAFNER: *The State of the Art in Ontology Design: A Survey and Comparative Review*, 1997

¹⁰⁶ WEIBEL et al.: *OCLC/NCSA Metadata Workshop Report*, 1995

¹⁰⁷ GRUBER: *A Translation Approach to Portable Ontology Specifications*, 1993

¹⁰⁸ FENSEL/GROENBOOM: *Specifying Knowledge-Based Systems with reusable Components*, 1997

¹⁰⁹ STUDER et al.: *Ontologies and the Configuration of Problem-solving Methods*, 1996

¹¹⁰ QUILLIAN: *Semantic Memory*, 1968

¹¹¹ MILLER et al.: *WordNet: An On-Line Lexical Database*, 1990

¹¹² INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: *ISO 13250: Information technology. SGML applications. Topic maps*, 2000

Standard XML Topic Maps¹¹³ dargestellt und ausgetauscht. Eine Themenkarte ist eine Sammlung von Themen oder Begriffen und semantisch typisierten Beziehungen zwischen diesen Themen. Diese Relationen, im Kontext von Themenkarten Rollen genannt, können zwei oder mehrstellig sein, sind aber nicht gerichtet. Externe Ressourcen als Individuen von Themenklassen werden durch Referenzen in Themenkarten eingebunden. Begrifflich klassifizierte Buchindizes oder Thesauren können so vorteilhaft dargestellt, sowie such- und visuell navigierbar gemacht werden. Semantische Netze beinhalten zwar dieselben strukturellen Elemente wie Ontologien (hierarchische Klassifikation von Individuen, typisierte Relationen und Attribute), entbehren jedoch einer formalen Semantik. Eine formale Semantik bedeutet die Spezifikation der Bedeutung einzelner Strukturelemente in ihrer Definition. Im Hinblick auf das Kriterium der Formalität, das heißt der maschinellen Verstehbarkeit, können semantische Netze daher nicht als Ontologien betrachtet werden.

Zur Formulierung einer Ontologie kommen spezielle Spezifikations Sprachen zum Einsatz. Neben der probabilistischen Darstellungen mit bayesschen Netzen¹¹⁴, welche auch objektorientiert erfolgen können¹¹⁵, sind traditionelle, prädikatenlogische Entwicklungen hierfür CYCL¹¹⁶, KIF¹¹⁷ (Knowledge Interchange Format) und, auf KIF beruhend, Ontolingua¹¹⁸. Auch die *Begrifflichen Graphen* (Conceptual Graphs) nach Sowa¹¹⁹, einer Erweiterung der semantischen Netze mit formaler Semantik, sind den Spezifikations Sprachen hinzu zurechnen. Auf diesen Sprachen aufbauend, wurden auch SHOE¹²⁰ (Simple HTML Ontology Extension), XOL¹²¹ (XML-Based Ontology Language) und OIL¹²² (Ontology Inference Layer) vorgestellt. Die zugrunde liegenden Repräsentationsparadigmen sind jedoch sehr unterschiedlich und beinhalten aus Gründen der endlichen Berechenbarkeit meist Untermengen der Prädikatenlogik erster Ordnung wie Beschreibungslogiken, Aussagelogiken und Frame-Logiken. In dieser Arbeit erfolgt keine Betrachtung der Darstellung von Ontologien mit bayesschen Netzen.

2.5.1.1 Aussagelogiken

Prinzipiell wird unter einer Aussage ein Satz verstanden, dem sinnvollerweise eines der Prädikate "wahr" oder "falsch" zukommt. Jede Aussage t_i besitzt einen *Wahrheitswert*. Wenn sie wahr ist, wird ihr die Zahl 1 als ihren Wahrheitswert zugeordnet, wenn sie falsch ist, die Zahl 0. Die Umgangssprache besitzt die Möglichkeit, aus gegebenen Aussagen neue Aussagen zu bilden. Das geschieht einmal durch Bildung der Verneinung einer Aussage und zum anderen durch die Verknüpfung zweier Aussagen t_1 und t_2 , durch die so genannten booleschen Junktoren \wedge (UND), \vee (ODER), \neg (NICHT) und dem Regel-Funktor \rightarrow (WENN, DANN). Diese Grundverknüpfungen lassen sich iterieren, so dass Aussagen beliebig großer Komplexität entstehen.

Zur exakten Beschreibung der Bedeutung der einzelnen Junktoren dienen Wahrheitstabellen. Die Aufstellung einer Wahrheitstafel erfolgt in Anlehnung an den umgangssprachlichen Gebrauch des jeweiligen Junktors. Im Falle der ODER-Verbindung und der WENN-DANN-Verknüpfung, deren umgangssprachliche Bedeutung ambivalent ist, entscheidet sich die Logik für eine der Bedeutungsvarianten¹²³.

Aus einer gegebenen Regelbasis D_0 lassen sich nun innerhalb eines Produktionsregelsystems durch Verkettung dieser Regeln durch den Modus Ponens automatisch Schlussfolgerungen ableiten. Dabei kommen zwei Verfahren zum Einsatz, die *Vorwärts-* und die *Rückwärtsverkettung*. Die Verfahren unterscheiden sich in der Verkettungsrichtung. Bei vorwärtsverkettenden Verfahren wird versucht, ausgehend von D_0 und einem bestimmten Term t_1 , die Menge aller

¹¹³<http://www.xtm.org>

¹¹⁴HELSPER/VAN DER GAAG: *Building Bayesian Networks through Ontologies*, 2002

¹¹⁵KOLLER/PFEFFER: *Object-Oriented Bayesian Networks*, 1997

¹¹⁶LENAT/GUAH: *Building large Knowledge-based Systems. Representation and Inference in the Cyc Project*, 1990

¹¹⁷GENESERETH/FIKES: *Knowledge Interchange Format Reference Manual - Version 3.0*, 1992

¹¹⁸GRUBER: *A Translation Approach to Portable Ontology Specifications*, 1993

¹¹⁹SOWA: *Conceptual Structures: Information Processing in Mind and Machine*, 1983

¹²⁰LUKE/HEFLIN: *SHOE 1.01. Proposed Specification*, 2000

¹²¹KARP/CHAUDHRI/THOMERE: *XOL: An XML-Based Ontology Exchange Language*, 1999

¹²²FENSEL et al.: *OIL in a Nutshell*, 2000

¹²³vgl. WOLKE: *Mathematik für Naturwissenschaftler*, 2004, S. 7ff

ableitbaren Terme t_i zu bestimmen, so dass $t_1 \rightarrow t_2$. Bei rückwärtsverkettenden Verfahren wird dagegen von einem zu bestimmenden Term t_i ausgegangen und versucht, diesen über Regeln auf die vorhandene Startdatenbasis D_0 auf t_1 zurückzuführen¹²⁴. Anschaulich bedeutet dies, dass bei Vorwärtsverkettung die Frage nach den Konsequenzen einer Aussage im Vordergrund steht, während bei Rückwärtsverkettung nach der Ursache einer Konsequenz gesucht wird.

Aussagelogiken in unterschiedlicher Ausdrucksfähigkeit sind Grundlage von formalen Sprachen wie PROLOG¹²⁵, CYCL, KIF und Ontolingua, welche wiederum auf KIF beruht. In diese Sprachenfamilie reihen sich auch framebasierte Formalismen wie Frame-Logik¹²⁶ (F-Logik) ein. F-Logik stellt im Wesentlichen ein auf dem objektorientierten Pradigma beruhendem PROLOG dar, das mit *Frames* angereicht wurde (vgl. Abbildung 2.2). *Frames*¹²⁷ sind hierbei als Beschreibungen von Objektklassen mit auch mehrwertigen Attributen, so genannten *Slots*, zu verstehen. Die Slots können an Objektinstanzen vererbt werden. F-Logik Ausdrücke können mittels der Lloyd-Topor Transformation¹²⁸ in Horn-Logik und damit in normale PROLOG-konforme Ausführungsanweisungen überführt werden.

	PROLOG	F-Logik
Logikdarstellung	$\text{mensch} \rightarrow \text{sterblich}$ $\text{sokrates} \rightarrow \text{mensch}$	$\forall x \in \text{mensch}[\text{hatEigenschaft} \rightarrow \text{sterblich}]$ $\text{sokrates} \in \text{mensch}$
Kodebeispiel	$\text{sterblich}(X) := \text{mensch}(X).$ $\text{mensch}(\text{sokrates}).$	$\text{mensch}[\text{hatEigenschaft} \Rightarrow \text{sterblich}].$ $\text{sokrates}:\text{mensch}.$
Schlussregel	Modus Ponens	Vererbung
Ergebnis	$\text{sokrates} \rightarrow \text{sterblich}$	$\text{sokrates}[\text{hatEigenschaft}] \rightarrow \text{sterblich}$

Tabelle 2.2: Beispiel "Alle Menschen sind sterblich" für Aussagelogiken

In ausdrucksstarken, regelbasierten Systemen ist es aufgrund der algorithmischen Entscheidbarkeit schwer, einer Regel $t_1 \rightarrow t_2$ nicht nur die Prädikate "wahr" oder "falsch", sondern auch Wahrscheinlichkeitswerte zuzuweisen, das heißt, mit welcher Wahrscheinlichkeit eine Regel im aktuellen Kontext gilt. Unter Einschränkung der Ausdrucksfähigkeit der jeweiligen Sprache ist dies jedoch möglich. So wurde von Fuhr eine probabilistische Erweiterung¹²⁹ von DATALOG eingeführt, bei der Regeln direkt gewichtet werden können. DATALOG ist eine sprachliche Einschränkung von PROLOG. Durch die Arbeiten von Koller und Pfeiffer¹³⁰ wurden auch framebasierte Systeme mit der Möglichkeit versehen, bestimmte Beziehungen und Attributierungen mit Wahrscheinlichkeitswerten zu versehen, wenn auch dies bei Regeln nicht möglich ist. Die komplette Aussagenbasis D_o wird dazu in ein bayessches Netz¹³¹ überführt, das mit konventionellen, in bayesschen Netzen üblichen, Inferenzverfahren ausgewertet wird.

2.5.1.2 Beschreibungslogiken

Beschreibungslogiken oder auch terminologische Logiken sind eine Familie von formalen Sprachen, welche im Gegensatz zur regelhaften Modellierung von Zusammenhängen, die Repräsentation und Verarbeitung der Semantik von natürlichsprachlichen Ausdrücken in den Vordergrund

¹²⁴vgl. BÜNING/LETTMANN: *Logik und Regelverarbeitung*, 2002, S. 155ff

¹²⁵CLOCKSIN/MELLISH: *Programming in Prolog*, 1987

¹²⁶KIFER/LAUSEN/WU: *Logical Foundations of Object-Oriented and Frame-Based Languages*, 1995

¹²⁷MINSKY: *A Framework for Representing Knowledge*, 1975

¹²⁸LLOYD/TOPOR: *Making Prolog more expressive*, 1984

¹²⁹FUHR: *Probabilistic Datalog - A Logic for powerful Retrieval Methods*, 1995

¹³⁰KOLLER/PFEFFER: *Probabilistic Frame-Based Systems*, 1998

¹³¹PEARL: *Probabilistic Reasoning in Intelligent Systems: Networks of plausible Inference*, 1988

stellen. Beschreibungslogiken sind historisch gesehen eine Weiterentwicklung der Aussagenlogiken, wobei sie den Vorteil einer eindeutigen und deklarativ definierten Semantik haben und nicht allzu mächtige Konstrukte, wie etwa Prozeduren, zulassen. Allerdings verwenden sie ähnliche Strukturen bei der automatischen Verarbeitung der Syntax von Ausdrücken, bei denen ebenfalls Klassen- und Instanzbeschreibungen mit Eigenschaften attribuiert werden können¹³².

Die zugrunde liegenden Ideen gehen dabei auf die Arbeiten von Brachman in der KL-ONE¹³³ (Knowledge Language One) genannten Sprache zurück. Dabei wird die, zur Beschreibung von Sachverhalten verwendete, Terminologie von der Beschreibung der Sachverhalte selbst getrennt. Dies entspricht weitgehend der in vielen Programmier- und Datenbanksprachen üblichen Trennung von Typ- und Objektebene. Die so genannte *terminologische* Komponente (T-Box), enthält deklaratives Wissen über Klassen von Individuen und die Beziehungen, die solche Klassen miteinander eingehen. Die *assertionale* Komponente (A-Box) enthält die Annahmen über die Individuen selbst.

Innerhalb der terminologischen Komponente kann durch Verwendung der bekannten Junktoren \wedge , \vee und \neg rekursiv aus *atomaren* Begriffen A komplexere Klassenbeschreibungen gebildet werden. Eine Klasse steht dabei für eine Menge von Individuen, die den Bedingungen der Klasse genügen. Zur algorithmischen Entscheidbarkeit wird zusätzlich der *universelle* Begriff \top (Top), sowie der *leere* Begriff \perp (Bottom) eingeführt.

In der Terminologie der Beschreibungssprachen werden sowohl die Beziehung zwischen Klassen als auch ihre Attribute als *Rollen* bezeichnet. Rollen sind das Äquivalent zu Slots, wobei nur zweiwertige Beziehungen erlaubt sind. In Beschreibungssprachen können Rollen mittels der Quantoren \exists , dem Existenzquantor, und \forall , dem Allquantor, definiert werden, sowie durch die Operatoren \leq (kleiner gleich) und \geq (größer gleich) in ihrer Anzahl restriktiert werden. Auch Rollen lassen sich rekursiv mittels der obigen Junktoren aus atomaren Rollen zu komplexeren synthetisieren.

Im Unterschied zur aussagelogischen Beschreibung werden Klassenhierarchien nicht nur durch explizite Angabe von Generalisierungsbeziehungen definiert, welche die Möglichkeit besitzen, legale Vererbungsinferenzen durch Überschreibung von Eigenschaftswerten zu unterbinden. Vielmehr wird zunächst festgelegt, welche Eigenschaften die Ausprägungen eine Klasse erfüllen müssen. Aufgrund solcher Beschreibungen kann dann abgeleitet werden, ob ein Begriff B einen anderen Begriff C subsumiert, das heißt, ob in Klasse C Klasse B enthalten ist. Formal wird die Aussage $C \subseteq B$ auf Wahrheit getestet. Dieser von Beschreibungslogiken bereitgestellte Inferenzdienst ermöglicht es Subsumptionsbeziehungen abzuleiten. Damit können sowohl dynamisch erzeugte Klassenbeschreibungen in einer Hierarchie korrekt eingeordnet als auch Individuen zu Klassen zugeordnet werden (siehe Abbildung 2.3).

Beschreibungslogik		
Logikdarstellung	$\text{mensch} = \text{sterblich} \sqcap \text{mensch}^*$ $\text{sokrates} \in \text{mensch}$	
Kodebeispiel	$\text{implies mensch and sterblich.}$ $\text{instance sokrates mensch.}$	
Schlussregel	Modus Ponens, Klassifikation	
Ergebnis	$\text{mensch} \subseteq \text{sterblich}$ $\text{sokrates} \in \text{sterblich}$	

Tabelle 2.3: Beispiel "Alle Menschen sind sterblich" für Beschreibungslogiken

Ausgehend von KL-ONE wurden eine Reihe von Beschreibungslogiken und zugehörige Sys-

¹³²vgl. NARDI/BRACHMAN: *An Introduction to Description Logics*, 2002, S. 7

¹³³BRACHMAN/SCHMOLZE: *An Overview of the KL-ONE Knowledge Representation System*, 1985

teme wie zum Beispiel CLASSIC¹³⁴, LOOM¹³⁵ oder in neuerer Zeit FACT¹³⁶ und RACER¹³⁷ vorgestellt. Auch diese unterscheiden sich im Wesentlichen durch den Umfang ihres Alphabets, das heißt komplexe Klassenbeschreibung aufgrund von Junktoren und Quantoren darzustellen. Auch die unterschiedlichen Restriktionen zur Verwendung des Nichtjunktors \neg kann zur Unterscheidung herangezogen werden.

Auch in diesem Bereich wurden probabilistische Erweiterungen geschaffen. Die Beschreibungslogik CLASSIC wird von Koller und Levy¹³⁸ ebenfalls auf Basis bayesscher Netze um die Angabe von Wahrscheinlichkeiten erweitert. Die expressivste Sprache mit Wahrscheinlichkeiten scheint zur Zeit P – SHOQ(D)¹³⁹ zu sein, mit der auch mit DAML+OIL (siehe Abschnitt 2.5.1.3) Ontologien geschlussfolgert werden kann.

2.5.1.3 Ontologiesprachen im weltweiten Netz

Aufgrund der Diversität der oben angeführten Sprachen versucht das World-Wide-Web-Konsortium (W3C) im praktischen Umfeld des Semantischen Netzes¹⁴⁰ Ontologiesprachen zu standardisieren. Die Basis dazu bildet das *Resource Description Frameworks*¹⁴¹ (RDF), eine Metadatenbeschreibung. Diese Bemühungen resultieren derzeit in den in Abbildung 2.4 dargestellten Sprachen. Die Schichtung der Ausdrucksfähigkeit von RDF zu RDF-Schema zu DAML+OIL und

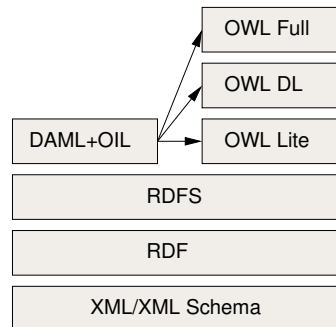


Abbildung 2.4: Ontologiesprachen auf Basis von RDF

der OWL Sprachfamilie folgt der Philosophie von Berners-Lee, in der ausdrucksfähigere Sprachen sich der Syntax und Semantik bestehender Standards bedienen¹⁴².

RDF

Genau wie XOL und OIL bedient sich RDF der *Extensible Markup Language* (XML) zur Persistenz und als Transportschicht. Dies hat den Vorteil einer syntaktischen Validierbarkeit mittels einer *Document Type Definition* (DTD). Zusätzlich erlaubt XML Schema die einheitliche Definition von Grunddatentypen wie Zeichenketten (Strings), natürliche und rationale Zahlen. RDF besitzt in etwa die Ausdrucksfähigkeit eines semantischen Netzes, das heißt, es können Begriffe und ihre Beziehungen untereinander spezifiziert werden. Sachverhalte werden in Ausdrücken, so genannten Subjekt-Prädikat-Objekt Tripeln, ausgedrückt. Übertragen auf Graphen stellen Subjekte und Objekte von Tripeln die Knoten eines Graphen und Prädikate die typisierten Kanten dar. Im

¹³⁴BORGIDA et al.: *CLASSIC: A Structural Data Model for Objects*, 1989

¹³⁵MACGREGOR: *Inside the LOOM Description Classifier*, 1991

¹³⁶HORROCKS: *The FaCT System*, 1998

¹³⁷HAARSLEV/MÖLLER: *RACER System Description*, 2001

¹³⁸KOLLER/LEVY/PFEFFER: *P-CLASSIC: A Tractable Probabilistic Description Logic*, 1997

¹³⁹GIUGNO/LUKASIEWICZ: *P-SHOQ(D): A Probabilistic Extension of SHOQ(D)*, 2002.

¹⁴⁰BERNERS-LEE: *Semantic Web Roadmap*, 1998

¹⁴¹Resource Description Framework, siehe <http://www.w3.org/RDF/>

¹⁴²vgl. HENDLER: *Agents and the Semantic Web*, 2001, Kap. 1

Gegensatz zu Themenkarten sind die Kanten daher gerichtet. Auch bietet RDF den Mechanismus der Reifikation, das heißt, es können Aussagen über Aussagen getroffen werden.

RDFS

Durch die Einführung einer modelltheoretischen, formalen Semantik kann RDF Schema¹⁴³ (RDFS) als basale Ontologiesprache der RDF Familie bezeichnet werden. Vordefinierte RDF Vokabeln erlauben die Konstruktion von Klassen und deren hierarchische Anordnung, sowie die Zuordnung von Individuen. Auch wird eine Trennung der Klassen- und Individuenebene eingeführt (vgl. A-Box, T-Box in Abschnitt 2.5.1.2), wobei allerdings eine Klasse Instanz einer anderen oder sogar derselben Klasse sein darf. Diese, Metamodellierung genannte, Konstruktion erlaubt die Überlagerung mehrerer semantischer Ebenen, allerdings auf Kosten vollständiger Inferenzmöglichkeiten in einer Ontologie¹⁴⁴. Inferenz kann nur auf einer auszuwählenden semantischen Ebene erfolgen, in welcher das Klassenmodell einer Diskurswelt eindeutig ist. Klassenbeziehungen sind ebenfalls Klassen und werden als Subklassen des vordefinierten Konstrukts *Eigenschaft* (Property) kenntlich gemacht. Durch die Angabe eines *Definitionsbereichs* (domain) und eines *Wertebereichs* (range) werden diese Beziehungen auf Klassen restriktiert. In der Terminologie der Beschreibungslogiken entspricht dies der Angabe eines Existenzquantors. Abbildung 2.5 stellt diese komplexen Zusammenhänge graphisch dar.

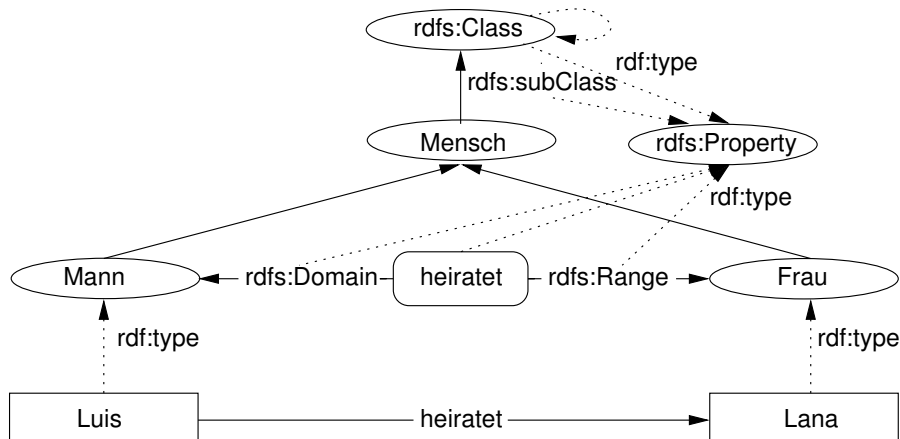


Abbildung 2.5: Beispielontologie in RDFS

DAML-OIL

Aufgrund großer sprachlicher Ähnlichkeit von OIL und der DARPA Agent Markup Language¹⁴⁵ (DAML) wurden die beiden Sprachen zu DAML-OIL verschmolzen, einem ersten Versuch zu einer standardisierten Ontologiesprache für das Semantische Netz. DAML+OIL ermöglicht eine ausdrucksstärkere Modellierbarkeit von Sachverhalten als RDFS und greift auf objektorientierte und framebasierte Sprachkonstrukte zurück. Im realen Umfeld ist es meist nötig, Klassen nicht nur durch Relationen in Beziehung zu setzen, sondern auch Attribute mit bestimmten Werten zu vergeben. Auch ist eine Spezifizierungsmöglichkeit der Kardinalität des Definitions- und Wertebereichs, sowie semantische Eigenschaften von Relationen von Nöten¹⁴⁶. In DAML+OIL wird daher zwischen Klassenrelationen (ObjectProperties) und Attributen (DatatypeProperties) unterschieden, die durch eine minimale und maximale Kardinalität restriktiert werden können.

¹⁴³BRICKLEY/GUHA: *RDF Vocabulary Description Language 1.0: RDF Schema*, 2003

¹⁴⁴vgl. WELTY/FERRUCCI: *What's in an Instance?*, 1994

¹⁴⁵HENDLER/MCGUINESS: *The DARPA agent markup language*, 2000

¹⁴⁶z.B. um aus zudrücken: Lana ist eine Frau von 30 Jahren und ist Ehefrau von genau einem Mann, nämlich Luis. "Ist Ehemann von" ist die inverse Relation zu "ist Ehefrau von".

Ebenfalls können Relationen als transitiv (*TransitiveProperty*), sowie invers (*InversePropertyOf*) oder gleich (*SamePropertyAs*) zu einer anderen deklariert werden. Zu einer komplexen Klassenkomposition kann noch die Vereinigung (*UnionOf*) beziehungsweise die Schnittmengenbildung (*IntersectionOf*) von Klassen herangezogen werden.

OWL

Ausgehend von DAML+OIL stellt schließlich OWL¹⁴⁷ (Web Ontology Language) den aktuellen Standard des W3C's zum Ausdruck ontologischer Information dar. OWL wurde um weitere Modellierungsprimitiven auf drei aufeinander aufbauende Detaillierungsebenen ausgedehnt, die einen Kompromiss zwischen Vollständigkeit und einfacher Nutzung darstellen:

- *OWL Lite* wurde als einfache Möglichkeit zur Klassifikation unter Verwendung elementarer Restriktionsmöglichkeiten konzipiert. Dieses soll sowohl die Entwicklung von Werkzeugen ohne aufwändigere Inferenzmechanismen unterstützen als auch eine schnelle Möglichkeit zum Umstieg von bereits bestehenden ontologischen Strukturen bieten. Der Hauptunterschied zu OWL DL besteht in der Angabe von Kardinalitätsrestriktionen, die nur zwischen 0 und 1 liegen dürfen.
- *OWL DL*¹⁴⁸ erlaubt formale Inferenz durch bestehende Beschreibungslogiksysteme und umfasst den vollen Sprachumfang. Allerdings wurden manche Sprachkonstrukte in ihrer Benutzung eingeschränkt, um eine algorithmische Vollständigkeit und Entscheidbarkeit innerhalb von Beschreibungslogiken zu gewährleisten. So bietet OWL Lite keine Möglichkeit zur Metamodellierung und es dürfen keine symmetrischen Relationen deklariert werden.
- *OWL Full* erlaubt Metamodellierung unter Verwendung des vollen Sprachumfangs auf Kosten der Vollständigkeit und Entscheidbarkeit. Es kann von sämtlichen RDF Konstrukten und der Reifikation Gebrauch gemacht werden.

Allen diesen hier vorgestellten Sprachen ist ihre implizite Axiomatik gemein. Zusätzliche Regeln in Form von WENN-DANN-Ausdrücken können hiermit nicht modelliert werden. In ihrem Aufbau ähneln sie daher den Beschreibungslogiken, wobei die Junktoren mit Hilfe von speziellen Relationen ausgedrückt werden.

2.5.1.4 Methoden und Verfahren zur Generierung von Ontologien

Ein wichtiger Schritt bei der Entwicklung eines wissensbasierten Systems ist, neben der tatsächlichen Nutzung der Wissensbasis, die Modellbildung der Anwendungsdomäne in Form einer Ontologie. Dabei handelt es sich um einen komplexen, iterativen und mehrstufigen Prozess, welcher nur teilweise automatisierbar ist. Die Mehrzahl der eingesetzten Methoden entsprechen bereits etablierten Techniken aus Bereichen der Informations- und Dokumentationswissenschaften, dem Software- und Datenbankentwurf, sowie der Linguistik und Lexikographie.

Da eine Ontologie definitionsgemäß erfordert, dass sich die Mitglieder ihrer Nutzergruppe auf die darin festgehaltenen Begriffsklärungen und -zusammenhänge einigen, ist es zunächst notwendig, einen Konsens unter den potenziellen Nutzern bezüglich der Inhalte zu erreichen. Dazu ist ein aufwändiger Erstellungsprozess erforderlich, der möglichst viele Nutzer einbezieht. Die von Uschold und Grüning¹⁴⁹ sowie von Guarino¹⁵⁰ vorgeschlagenen Methodiken unterscheiden bei der Entwicklung und Pflege von Ontologien im Hauptsächlichen die folgenden vier Phasen:

1. Problemanalyse und Anforderungsspezifikation
2. Modell- und Begriffsbildung

¹⁴⁷ MCGUINNESS/VAN HARMELLEN: *OWL Web Ontology Language Overview*, 2003

¹⁴⁸ DL = Description Logic

¹⁴⁹ USCHOLD/GRÜNINGER: *Ontologies: Principles, Methods, and Applications*, 1996

¹⁵⁰ GUARINO: *Formal Ontology, Conceptual Analysis and Knowledge Representation*, 1995

3. Formalisierung und Axiomatisierung

4. Evaluierung

Ausgehend von einer konkreten Aufgabe wird im *Analyse- und Anforderungsschritt* die Einsatzdomäne der Ontologie abgegrenzt. Dies beinhaltet die Erkennung von typischen Nutzungsszenarien, sowie Aufgabenstellungen, die mit Hilfe der Ontologie gelöst werden sollen. Dazu werden informell so genannte Befähigungsfragen gestellt, die von ihr zu beantwortet sind.

Die *Modell- und Begriffsbildung* erfolgt unter Befragung von potentiellen Nutzern und Domänenexperten, sowie der Auflistung von Schlüsselbegriffen und deren Beziehungen untereinander innerhalb der Anwendungsdomäne. Kooperativ erfolgt nun eine Übereinkunft zur zunächst natürlichsprachlichen Definition dieser Begriffe und Relationen. Eine hierarchische Anordnung dieser Begriffe kann entweder von oben nach unten oder von unten nach oben erfolgen. Auch eine Mischform ist einsetzbar, in welcher von wichtigen Begriffen ausgegangen wird und diese verallgemeinert oder spezialisiert werden.

Im *Formalisierungs- und Axiomatisierungsschritt* werden die natürlichsprachlichen Begriffsdefinitionen in einer Ontologiesprache ausgedrückt. Begriffe werden solange rekursiv aus anderen Begriffen und Axiomen definiert, bis nur noch primitive Begriffe übrig bleiben, welche nicht mehr weiter vereinfacht werden können. Die Axiome müssen die Bedeutung der Begriffe klar umreißen in dem Attribute, Randbedingungen und Beziehungen ausgedrückt werden.

Die so entstandene Ontologie wird in der *Evaluierungsphase* anhand der Befähigungsfragen auf Vollständigkeit getestet. Außerdem müssen Kriterien zur Konsistenzprüfung gefunden werden. Können nicht alle Befähigungsfragen beantwortet werden oder existieren widersprüchliche Definitionen, so muss die Ontologie erweitert oder korrigiert werden. Zur Unterstützung dienen Autorenwerkzeuge wie Protege 2000¹⁵¹, OntoEdit¹⁵² oder OilEd¹⁵³. Protege 2000 verfügt über eine inzwischen reichhaltige Landschaft von modularen Erweiterungen zur Erstellung und Visualisierung von Ontologien, während OntoEdit und OilEd durch integrierte Inferenzmechanismen die Evaluation von Ontologien erleichtern. Auch können gängige Sprachen wie OWL und DAML+OIL bearbeitet werden.

Ausgehend von dem in IEEE 1074-1995¹⁵⁴ standardisierten Softwareentwicklungsprozess definiert die als Methontology¹⁵⁵ bekannte Methodik den Ontologieentwicklungsprozess (Ontology Development Process - ODP). Die zu entwickelnde Ontologie wird hierin als evolutionärer Prototyp aufgefasst, der die obigen Phasen in einem Lebenszyklus durchläuft und iterativ verfeinert wird. Zusätzlich werden begleitende Aktivitäten wie Projekt- und Konfigurationsmanagement erkannt. Eine detaillierte Übersicht hierzu und Vergleiche zwischen weiteren, sehr ähnlichen Methodiken liefern die Artikel von Jones et al.¹⁵⁶ und Fernandez-Lopez¹⁵⁷.

An Ontologien werden besondere Ansprüche hinsichtlich ihrer Konsistenz und Qualität gestellt. Die Modellierung von Ontologien kann daher nur semiautomatisch erfolgen, indem Begriffs- und Modellvorschläge von Textanalysewerkzeugen aus domänenspezifischen Dokumenten generiert werden, die dann intellektueller Überarbeitung durch menschliche Editoren unterworfen werden. Abbildung 2.6 zeigt eine Verallgemeinerung des von Maedche und Staab¹⁵⁸ vorgeschlagenen Rahmenwerks Text-To-Onto zur Extraktion von Ontologien aus natürlichsprachlichen Texten. Ähnliche Ansätze existieren auch innerhalb des LTG Projekts der Universität Edinburgh¹⁵⁹ und des Systems ASIUM¹⁶⁰. Verschiedenste Techniken zur Verarbeitung der natürlichen Sprache

¹⁵¹ NOY et al.: *Creating Semantic Web Contents with Protege-2000*, 2001

¹⁵² STAAB/MAEDCHE: *Ontology Engineering beyond the Modeling of Concepts and Relations*, 2000

¹⁵³ BECHHOFFER et al.: *OilEd: a Reason-able Ontology Editor for the Semantic Web*, 2001

¹⁵⁴ IEEE COMPUTER SOCIETY: *IEEE 1074-1995: Standard for Developing Software Lifecycle Processes*, 1995

¹⁵⁵ FERNANDEZ-LOPEZ/GOMEZ-PEREZ/JURISTO: *METHONTOLOGY: From Ontological Art Towards Ontological Engineering*, 1997

¹⁵⁶ JONES/BENCH-CAPON/VISSER: *Methodologies for Ontology Development*, 1998

¹⁵⁷ FERNANDEZ-LOPEZ/GOMEZ-PEREZ/JURISTO: *METHONTOLOGY: From Ontological Art Towards Ontological Engineering*, 1997

¹⁵⁸ MAEDCHE/STAAB: *Mining Ontologies from Text*, 2000

¹⁵⁹ MIKHEEV/FINCH: *A Workbench for Finding Structure in Texts*, 1997

¹⁶⁰ FAURE/NÉDELLEC/ROUVEIROL: *Acquisition of Semantic Knowledge using Machine Learning Methods: The System ASIUM*, 1998

und des maschinellen Lernens werden dabei integriert, um den Modellierer bei der Identifikation von relevanten Begrifflichkeiten und deren Beziehungen zu unterstützen.

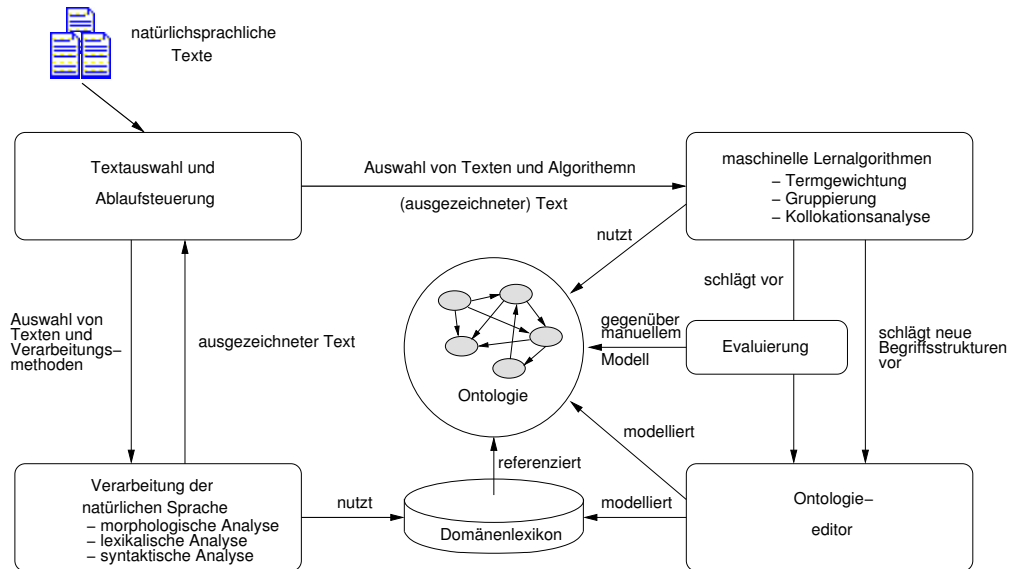


Abbildung 2.6: Rahmenwerk zur semiautomatischen Erstellung von Ontologien

Nach Auswahl eines domänenspezifischen Dokumentenkörpers erfolgt eine Behandlung der Texte mit Techniken zur Verarbeitung der natürlichen Sprache. Diese Texte werden mit der gewonnenen linguistischen Information ausgezeichnet. Die Analyse erkennt Wortformen und -typen und löst deren unterschiedliche Bedeutungen anhand des lokalen Kontexts innerhalb eines Satzes auf. Wortformen werden in ein Domänenlexikon eingetragen, welches eine Abbildung dieser Wörter auf die Begriffe der Ontologie ermöglicht. Namen und Mehrwortbegriffe stellen Kandidaten der Begriffe innerhalb der Ontologie dar. Durch eine syntaktische Analyse werden Abhängigkeitsrelationen zwischen den Subjekten und Objekten der Sätze extrahiert. Diese werden als Indikatoren für mögliche semantische Beziehungen der Subjekt-Objekt Paare aufgefasst.

Begriffskandidaten und Relationsvorschläge können in die Ontologie übernommen werden und dienen als Hintergrundwissen für maschinelle Lernverfahren. So können relevante Begriffe mittels Termgewichtung (siehe Abschnitt 2.4.2) identifiziert werden. Vorschläge zur Klassifikation der ontologischen Begriffe werden durch Gruppierungsverfahren erstellt, in dem zum Beispiel angenommen wird, häufige Begriffe wären allgemeiner als seltenere. Auf Basis dieser Taxonomie können nichthierarchische Beziehungen durch Kollokationsanalyse unter Berücksichtigung der Subjekt-Objekt Paare gewonnen werden. Bei der Kollokationsanalyse werden Wörter aufgrund der statistischen Häufigkeit ihres gemeinsamen Auftretens innerhalb eines Satzes durch so genannte Assoziationsregeln in Beziehung gesetzt. Assoziationsregeln¹⁶¹ wurden ursprünglich innerhalb der Warenkorbanalyse zur Bestimmung von Kundenkaufverhalten entwickelt.

Die hier verwendeten maschinellen Lernverfahren können auch ohne linguistisches Hintergrundwissen auf rein statistischer Basis angewandt werden. Allerdings werden dann sehr große Korpora zur Erzielung guter Ergebnisse benötigt. Ein Beispiel hierfür wäre das Wortschatzprojekt¹⁶² der Universität Leipzig zur Erstellung eines deutschen Thesaurus.

¹⁶¹ AGRAWAL/MIELINSKI/SWAMI: *Mining Association Rules between Sets of Items in large Databases*, 1993

¹⁶² HEYER et al.: *Learning Relations using Collocations*, 2001

2.5.2 Verarbeitung der natürlichen Sprache

Automatische Verarbeitung der natürlichen Sprache (Natural Language Processing - NLP) ist eine Sammlung von Techniken zur Analyse und Repräsentation natürlicher Texte auf einer oder mehrerer Ebenen der linguistischen Analyse. Die Zielsetzung ist dabei, eine menschenähnliche Verarbeitung der Sprache für eine Reihe bestimmter Aufgaben oder Anwendungen wie Information Retrieval, maschinelles Übersetzen und oder in Dialogsystemen zu erreichen. Dabei werden die nachfolgend beschriebenen Prozessschritte innerhalb der linguistischen Analyse unterschieden¹⁶³, die auch unterschiedliche Bedeutung für das IR besitzen¹⁶⁴.

Morphologisch-lexikalische Analyse

In der *morphologisch-lexikalische Analyse* wird bestimmt, wie Wörter durch Flexion, Derivation und Komposition aus weit einfacheren, Morpheme genannten, Begriffseinheiten zusammengesetzt sind. Flexion bezeichnet die Variation, mit der sich ein Wort an verschiedene syntaktische Umgebungen anpasst, beziehungsweise verschiedene syntaktische und semantische Funktionen ausübt (z.B. Buch, Buch/es, lern/en, lern/e, lern/te). Unter Derivation wird die Verbindung eines Wortes mit einem Prä- oder Suffix verstanden, woraus ein neues Wort entsteht (z.B. Schön/heit, er/lernen, unter/schied/lich). Durch Komposition werden zwei oder mehrerer Wörter zu einem neuen Wort verknüpft (z.B. Haus/schuh, gras/grün). In diesem Schritt werden Wörter auf ihre Stamm- oder Grundform zurückgeführt, sowie ihre Wortkategorie (Nomen, Verb, etc.) und Eigenschaften bestimmt. Für diese Analyse wird normalerweise ein Lexikon verwendet, welches die nötige Information über die Wörter enthält. Das Lexikon muss auch Informationen für die Syntax- und Semantikkomponenten der höheren Analyseebenen enthalten. Es existieren allerdings auch regelbasierte Algorithmen zur Stammformreduktion wie der von Porter¹⁶⁵, die ohne Lexikon auskommen.

Die Wortstammreduktion hat sowohl bei der Anfrage als auch im Indexierungsschritt lange Tradition im IR. Mit ihr kann die Vollständigkeit verbessert werden, da zum Beispiel auch Pluralformen erkannt werden, wenn nur die Singularform eines Wortes als Suchbegriff verwendet wurde. Ebenfalls kann der zu speichernde Dokumentindex verkleinert werden, wenn nur Grundformen der Indexterme abgelegt werden.

Syntaktische Analyse

Die *syntaktische Analyse* befasst sich mit der Rolle des einzelnen Wortes und dessen Beziehung zu den anderen Worten innerhalb eines Satzes anhand grammatikalischer Regeln. So werden unter anderem Subjekt, Prädikat und Objekt des Satzes identifiziert. Für die Syntaxanalyse werden Grammatiken verwendet, die aus rekursiven Regeln bestehen. Sie werden auch generative Grammatiken genannt, weil durch wiederholtes (unendliches) Anwenden der Grammatikregeln eine Menge von Sätzen generiert werden kann. Die von der Grammatik generierte Satzmenge sollte im Idealfall genau gleich der einer natürlichen Sprache sein. Es existieren jedoch derzeit noch keine generativen Grammatiken, welche eine natürliche Sprache vollständig erzeugen, da die Konstruktion einer solchen Grammatik sehr aufwändig wäre.

Im IR erlaubt die syntaktische Analyse vor allem eine Phrasenerkennung in der Suchanfrage und im Dokument. Gerade Phrasen wie "goldene Mitte" eignen sich als gute Such- und Indexterme, da sie meist mehr Bedeutung tragen als einzelne Wörter.

Semantische Analyse

In der *semantischen Analyse* werden sowohl die unterschiedlichen Bedeutungen eines Wortes abgeleitet, als auch die Bedeutungen, die sich für ein Wort aus dem Satzzusammenhang ergeben. Hierbei wird aus den syntaktisch analysierten Sätzen eine interne Repräsentation der

¹⁶³vgl. ALLEN: *Natural Language Understanding*, 1995, S. 6

¹⁶⁴vgl. LIDDY: *Enhanced Text Retrieval Using Natural Language Processing*, 1998, S.2

¹⁶⁵PORTER: *An Algorithm for Suffix Stripping*, 1980

Satzbedeutung abgeleitet. Für diese Repräsentation wird meist eine Art von Logik verwendet. Die Prädikatenlogik erster Ordnung ist dafür allerdings nur bedingt geeignet, da natürliche Sprachen viele Ungenauigkeiten enthalten, welche durch die Prädikatenlogik nicht richtig dargestellt werden können. Die Prädikatenlogik wurde daher auf verschiedene Arten erweitert. Diese Erweiterungen versuchen auch ungenaue Ausdrücke wie "vielleicht", "wahrscheinlich", etc. zu modellieren.

Eine der wichtigsten Annahmen, die dazu oft gemacht wird, ist die Betrachtung der semantischen Analyse als einen kompositionellen Prozess. Die Bedeutung eines komplexen Ausdrucks ist hierbei eine Funktion der Bedeutung seiner Teilausdrücke und der Art ihrer Zusammensetzung¹⁶⁶. Die kleinste Bedeutungseinheit bildet dabei ein einzelnes Wort. Diese Bedeutungseinheiten können dann theoretisch einfach nach der aus der Syntaxisanalyse erhaltenen Struktur zusammengesetzt werden. Allerdings sind solche kompositionellen Semantiktheorien schwierig zu entwickeln, da sich die Strukturen der semantischen Repräsentation und der Syntaxisanalyse nicht immer gleichen.

Der semantische Analyseschritt erlaubt die Begriffsklärung von Wörtern mit unterschiedlicher Bedeutung und die Identifikation von Prädikaten. Dadurch kann eine Anfrage mit zusätzlichen Termen erweitert werden, die in einem semantischen Zusammenhang mit der Anfrage stehen. Dies unterstützt die Anfrageformulierung, da zusätzliche relevante Dokumente gefunden werden, in denen die Suchterme nicht direkt auftreten.

Pragmatische Analyse

Die *pragmatische Analyse* setzt das Ergebnis der semantischen Analyse in den Kontext und versucht so die endgültige Bedeutung des Satzes als interne Repräsentation zu modellieren. Der Kontext kann dabei in den Textkontext und den Situationskontext aufgeteilt werden. Die Betrachtung des Textkontext wird auch als *Diskursanalyse* bezeichnet. Der Textkontext besteht aus einem lokalen Teil mit detaillierten Informationen über den vorangegangenen Satz und aus einem globalen Teil über den ganzen Text, der die besprochenen Themen beinhaltet. Der Situationskontext besteht einerseits aus Informationen über die Umgebung wie etwa Ort, Zeit und sichtbare Objekte und andererseits aus dem Hintergrund- oder Weltwissen des Sprechers und Hörers beziehungsweise des Computers. In dem pragmatischen Analyseschritt gilt es, Querreferenzen in Sätzen aufzulösen. So werden Pronomen durch die von ihnen referierten Objekte im Textkontext ersetzt, während Nomen mit Referenzen auf Objekte aus dem Situationskontext in Beziehung gesetzt werden. Hierzu wird detailliertes Weltwissen benötigt, was an folgendem Beispiel verdeutlicht wird: Im Satz "Der Vogel setzte sich auf den Ast. Kurz darauf flog er davon" bezieht sich das Pronomen "er" auf den "Vogel", da nur Vögel und in der Regel keine Äste davonfliegen. In einem ähnlichen Satz "Der Vogel setzte sich auf den Ast. Kurz darauf brach er ab" bezieht sich "er" jedoch auf den Ast, da Vögel nicht abbrechen. Ebenfalls wird für das Auflösen von Querbeziehungen Hintergrundwissen benötigt. In "Er nahm die Zeitung und fing an zu lesen" findet sich die implizite Annahme, er liest die Zeitung.

Im IR ermöglicht die Diskursanalyse die Rolle von Textabschnitten in Dokumenten zu bestimmen. Ob in einem Textabschnitt Meinungen, Schlussfolgerungen oder Fakten dargestellt werden, kann für die Bestimmung der Relevanz auf eine Suchanfrage herangezogen werden. Zusätzlich ist ein Verständnis des Situationskontextes im Dialog des IR-Systems mit dem Nutzer von besonderer Bedeutung für die Bestimmung und Erfüllung des tatsächlichen Informationsbedürfnisses.

Alle diese Ebenen der linguistischen Analyse greifen ineinander und befassen sich, von oben nach unten gesehen, mit immer größeren Analyseeinheiten¹⁶⁷. Je größer diese Analyseeinheiten werden, also vom Morphem über das Wort, den Satz, den Abschnitt zum gesamten Dokument, desto weniger präzise werden die anwendbaren Modelle und Verfahren, da immer weniger Regeln erkennbar werden und immer mehr Hintergrundwissen benötigt wird. Höhere Ebenen setzen verstärkt das Sprachverstehen der unteren Ebenen voraus und die eingesetzten Theorien zur Erklärung der Daten kommen immer häufiger aus den Bereichen der kognitiven Psychologie und

¹⁶⁶auch Fregesches Prinzip, benannt nach dem Mathematiker und Philosophen Gottlob Frege (1848- 1925)

¹⁶⁷vgl. FELDMAN: *NLP meets the Jabberwocky - Natural Language Processing in Information Retrieval*, 1999, S. 13

der künstlichen Intelligenz. Demzufolge wurden bisher hauptsächlich die unteren Ebenen der Sprachverarbeitung gründlicher untersucht und in bestehende IR-Systeme integriert¹⁶⁸.

2.5.3 Logische Ansätze

"Logikbasierte Ansätze zum Information Retrieval machen die Evaluation einer Anfrage zu einem Prozess der Schlussfolgerung auf Grundlage von semantischer Beziehungen"¹⁶⁹. Voraussetzung hierfür ist die Repräsentation von Dokumenten und Anfragen mittels logischer Regeln. Anfragen und Dokumente können dann durch einen Inferenzprozess zueinander in Beziehung gesetzt werden. Ein Dokument d wird als relevant zu einer Anfrage q angesehen, wenn das IR-System die Anfrage aus denjenigen Regeln und Aussagen ableiten kann, welche das Dokument d beschreiben. Dabei kann der Ableitungsprozess prinzipiell komplizierter sein als lediglich die Überprüfung des Vorkommens von gleichen Termen in Dokument und Anfrage. Insbesondere kann zusätzliches, nicht aus Dokumenten extrahiertes, Wissen in Form von Regeln und Aussagen verwendet werden. Dieses Wissen kann aus einer Wissensbasis des Systems stammen oder von den Nutzenden eingegeben beziehungsweise abgefragt werden. Die Anfragebearbeitung entspricht damit der herkömmlichen Logik von Datenbanken, bei der ebenfalls alle Objekte o zurück geliefert werden, welche die Implikation $o \rightarrow q$ als wahr erfüllen¹⁷⁰.

Dieses Datenbankmodell wurde von van Rijsbergen auf den Anwendungsbereich des Information Retrieval übertragen¹⁷¹¹⁷². Mit Einbeziehung der im Information Retrieval immanenten Ungenauigkeit und Vagheit wird dabei die Retrieval-Funktion R im Sinne einer logischen Formel

$$R_l := P(d \rightarrow q) \quad (2.11)$$

verstanden. Der Operator " \rightarrow " stellt hierbei den bedingten Wahrheitswertfunktork dar, der durch eine geeignete Logik formalisiert wird. $P(\alpha)$ ist als die Wahrscheinlichkeit anzusehen, dass die Aussage $d \rightarrow q$ als gültig einzuschätzen ist.

"Wahrscheinlichkeit hat hier eine Schlüsselrolle inne"¹⁷³, da sich logische Ausdrücke nur bedingt zur Repräsentation von Dokumenten und Anfragen eignen, die ein menschliches Informationsbedürfnis ausdrücken. Dokumente können daher nur mit einem bestimmten Grad an Sicherheit als relevant zu einer Anfrage eingestuft werden. Zur Berechnung dieser Wahrscheinlichkeit bedient sich van Rijsbergen in seinem Modell einem *Imaging* genannten Verfahren, welches sich aber im Wesentlichen auf das Vektorraummodell reduzieren lässt¹⁷⁴.

Vor allem aber ist in diesem Zusammenhang die modelltheoretische, semantische Ebene von Logik interessant. Relevanz ist darin mit dem mengenmäßigen Informationsgehalt beziehungsweise mit Bedeutung deckungsgleich und es werden nur die Dokumente zurück geliefert, deren Inhalt mit dem einer Anfrage übereinstimmen. Die Entwicklung von effektiven IR-Systemen bedeutet daher, formale Theorien der Semantik von Dokumenten zu bilden, wobei kontext- und situationsbezogene Aspekte von Information einbezogen werden können. Aber auch der beweistheoretische Charakter der Logik ist von Bedeutung. Aufgrund solcher Repräsentation von Anfragen und Dokumenten können abstrakte Experimente durchgeführt werden, die anhand eines automatischen Theorembeweisera heraus zu finden vermögen, ob die Aussage $d \rightarrow q$ gilt.

Im logischen IR kann entweder der gesamte Informationsraum oder jedes einzelne Dokument als mögliche Welt modelliert werden. Im ersten Fall kann der Wahrheitsstatus von $d \rightarrow q$ direkt gefolgert werden. Im Falle der Betrachtung von Dokumenten als jeweils mögliche Welt, wird nach denjenigen Welten gesucht, innerhalb denen $d \rightarrow q$ gilt. Diese prinzipielle Unterscheidung betrifft vor allem die Komplexität des jeweiligen Modells, welches im Rahmen der automatischen Schlussfolgerung keine Widersprüche enthalten darf. Werden Dokumente jedoch als mögliche

¹⁶⁸vgl. LIDDY: *Enhanced Text Retrieval Using Natural Language Processing*, 1998, S. 3

¹⁶⁹NIE: *Towards a Probabilistic Modal Logic for Semantic-Based Information Retrieval*, 1992

¹⁷⁰vgl. VOSSEN: *Datenbankenmodelle, Datenbanksprachen und Datenbankmanagement*, 1999

¹⁷¹VAN RIJSBERGEN: *A new theoretical framework for Information Retrieval*, 1986a

¹⁷²VAN RIJSBERGEN: *A non-classical Logic for Information Retrieval*, 1986b

¹⁷³übersetzt aus: SEBASTIANI: *A Note on Logic and Information Retrieval*, 1996, S. 6

¹⁷⁴vgl. FERBER: *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*, 2003, S. 125

Welten betrachtet, so können die unterschiedlichen Welten durchaus widersprüchliche Aussagen enthalten. Dies ist auch der Grundgedanke des Semantischen Netzes. Das herkömmliche Netz, bestehend aus beliebig komplexen Seiten, wird durch maschinenversteh- und interpretierbare, semantischen Daten, durch Ontologien annotiert. Jede Ontologie ist dabei eine mögliche Welt, welche darauf überprüft wird, ob eine gegebene Anfrage q gefolgert werden kann.

Neben der Darstellung von Inhalten durch logische Aussagen gilt es ebenfalls, denjenigen Implikationsfunktoren beziehungsweise diejenige Logik \mathcal{L} herauszufinden, deren Implikationsfunktoren Relevanz am Besten darstellen¹⁷⁵. Welche hierbei die geeignetste ist, ist bis heute aufgrund von Schwierigkeiten in ihrer Ausdruckskraft und ihres Aufwands beziehungsweise Komplexität der Berechnungen noch nicht geklärt. Die Auswahl hängt dabei vom konkreten Anwendungsfall ab.

2.5.4 Diskussion

Die Qualität eines IR-Systems lässt sich im Allgemeinen durch die Einbeziehung von Hintergrundwissen verbessern. Verfahren zur Verarbeitung der natürlichen Sprache verbessern das maschinelle Verständnis von Anfragen und Dokumenten. In diesem Sinne hat zum Beispiel die Wortstammreduktion innerhalb der morphologischen Analyse eine lange Tradition im IR, sowohl bei der Anfrage als auch im Indexierungsschritt. Rein regelbasierte Algorithmen zur Stammformreduktion neigen jedoch zur Erzeugung nicht existenter Stämme (z.B. Information -> Inform) und künstlicher Ambiguitäten (z.B. Herr, herrlich -> herr). Auch lassen sich für flexionsreiche Sprachen, wie dem Deutschen, nur unvollständige Algorithmen finden. Eine vollständige morphologisch-lexikalische Analyse ermöglicht jedoch die Identifikation aller Wortformen und Mehrwortbegriffe. Eine Suche nach "Vertragshandys" würde auch "Handy mit Vertrag" finden. Neben dem Effekt der Datenreduktion des Dokumentenindexes wird dabei eine Verbesserung der Vollständigkeit ohne Verschlechterung der Genauigkeit angestrebt. Inhaltlich gesehen stellt das eine Verallgemeinerung der Bedeutung des Wortes dar. Mit dem grundformreduzierten Term sollten daher mehr Dokumente zu einem Thema gefunden werden. Obwohl diese Annahme in der Literatur umstritten ist¹⁷⁶, zeigen neuere Studien¹⁷⁷, dass die Retrievalqualität für germanische Sprachen mit einer sorgfältigen Grundformreduktion und einer guten Dekomposition von Mehrwortbegriffen deutlich erhöht wird.

Die höheren Ebenen der Sprachverarbeitung wie die semantische und pragmatische Analyse versprechen weitere qualitative Verbesserungen. Zum einen liefern sie sowohl logische Modelle der Semantik der analysierten Texte als auch Beweise für die Richtigkeit der Suchantworten. Zum anderen kann im Dialog mit dem IR-System der Informationsbedarf des Suchenden präzisiert werden. Neben dem Fehlen einer generativen Grammatik, die alle Phänomene der natürlichen Sprachen abdeckt, ist das Hauptproblem hierbei jedoch die Auflösung von Mehrdeutigkeiten. Diese könnten theoretisch durch Einbeziehung des Kontextes und gezieltes Nachfragen aufgelöst werden. Dazu wäre aber menschliches Weltwissen nötig. Die symbolbasierte maschinelle Darstellung von Weltwissen leidet an den klassischen Schwierigkeiten der künstlichen Intelligenz der siebziger Jahre, während konnektionistische Methoden keine expliziten Sachverhalte auszudrücken vermögen. Außerdem fehlen empirische Untersuchungen, die den Beitrag zur Qualität der Suche der heutigen Sprachanalyse hervorheben¹⁷⁸.

Logikbasierte Ansätze zum Information Retrieval besitzen, neben ihrem modell- und beweistheoretischem Charme, die Möglichkeit, Sachverhalte unabhängig von ihrer sprachlichen Ausdrucksform und maschinenverstehbar darzustellen. Der Ansatz des Semantischen Netzes begegnet der Unmöglichkeit der Modellierung der Welt im Sinne der KI mit Hilfe verteilter Ontologien, die immer nur einen kleinen Teilausschnitt der Welt, nämlich Webseiten, darstellen. Auf eine Suchanfrage wird dann die Ontologie oder Webseite gefunden, in der die Suchanfrage gefolgert werden kann. Die ontologische Modellierbarkeit von Texten ist begrenzt, da "eine

¹⁷⁵vgl. SEBASTIANI: *A Note on Logic and Information Retrieval*, 1996, S. 6f

¹⁷⁶vgl. FERBER: *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*, 2003, S. 23

¹⁷⁷BRASCHLER/RIPPLINGER: *How Effective is Stemming and Decomposition for German Text Retrieval?*, 2004

¹⁷⁸vgl. LIDDY: *Enhanced Text Retrieval Using Natural Language Processing*, 1998, S.2

vollständige Analyse der natürlichen Sprachen im Rahmen der logischen Semantik prinzipiell unmöglich ist¹⁷⁹. Zum Beispiel führen Sätze, die abgewandelte Formen des Epimenides- oder Lügner-Paradoxons¹⁸⁰ beinhalten, bei der logischen Analyse zu einem Widerspruch. Eine weitere Schwierigkeit ergibt sich aus dem so genannten "Problem des völlig uninteressanten Dokuments (VUD)"¹⁸¹. Ein VUD trägt keine Information und lässt sich in einer Sprache mit drei Wörtern $D = \{d_1, d_2, d_3\}$ durch $\neg d_1, \neg d_2, \neg d_3$ ausdrücken. Die Aussage $d \rightarrow q$ wäre dann jedoch für jede Anfrage q wahr.

Aufgrund dieser Widrigkeiten ist der nahe liegende Ansatz, Dokumente nicht direkt zu modellieren, sondern eine Kombination von Domänenontologien und merkmalsbasierten IR-Techniken einzusetzen. Studer identifiziert dabei vier wesentliche Vorteile¹⁸²:

1. *Verfeinerte Suchunterstützung*: der Zugriff über eine Schlüsselwortbasierte Suche wird über die Inferenzmaschine und die ihr zugrunde liegende Modellierung der Ontologie erweitert und verfeinert. Begriffe können über die Ontologie automatisch mit Synonymen und verknüpften Begriffen erweitert werden. Die Suche wird dadurch intuitiver.
2. *Navigationsunterstützung im Kontext*: ein zusätzlicher Nutzen entsteht durch die Navigationsmöglichkeiten der Ontologie. Der Nutzer kann Verfeinerungs- oder Verallgemeinerungsschritte durchführen; er wird durch seinen Suchkontext geleitet.
3. *Auswertung der Zusammenhänge*: ein weiterer Zusatznutzen entsteht durch die Auswertung der regelbasierten Zusammenhänge mittels Inferenzmaschinen. Das implizite Wissen kann somit ebenfalls abgefragt und dargestellt werden.
4. *Einfache Integration strukturierter Informationsquellen*: durch einfache Verknüpfungsmöglichkeiten der Ontologie mit Datenbankschemata können strukturierte Informationsquellen parallel mit Schlüsselwort-basierten Suchanfragen ausgewertet und gemeinsam als Antworten dargestellt werden.

Diese Vorteile stehen prototypisch für die Erwartungen an die Technologien des Semantischen Netzes, sind jedoch nicht unproblematisch. Im ersten Punkt wird die Ontologie als Thesaurus für eine semantische Analyse der Anfrage eingesetzt. Dadurch lässt sich die Anfrage mit Begriffen erweitern, die in einem bedeutungsvollen Zusammenhang zu den Anfragetermen stehen. So werden auch Umgebungstreffer in die Ergebnismenge mit einbezogen. In der Literatur zeigt sich jedoch, dass diese zusätzliche Unschärfe im Allgemeinen keine Verbesserung der Suchqualität erbringt¹⁸³. Vorteile können allenfalls bei kleinen Dokumentensammlungen erzielt werden, wenn auf eine Suchanfrage keine direkten Treffer gefunden werden konnten. In diesem Falle sind Umgebungstreffer erwünscht.

Durch die Augmentierung traditioneller IR-Techniken mit Ontologien und der Integration von strukturierten Datenquellen wird zusätzliche Komplexität in der Schnittstelle zwischen Mensch und Maschine eingeführt. Wird die Ontologie direkt abgefragt, um aus den Regeln zusätzliche Information zu gewinnen, muss sich dazu einer logischen Sprache bedient werden. Deren Formulierung ist für den Standardnutzer nicht unbedingt intuitiv. Auch muss die aus der Ontologie gewonnene Information, sowie Treffer der konventionellen Suchmaschine in die Nutzerschnittstelle integriert werden. Auf die Gestaltung einer Nutzerschnittstelle ist daher ein besonderes Augenmerk zu richten, damit die aufgezeigten Vorteile einen tatsächlichen Nutzen erzielen.

Abschließend sei bemerkt, dass diesen Vorteilen die gezeigte Aufwändigkeit und die damit verbundenen hohen Kosten zur Erstellung der Ontologien gegenüber stehen. Auch ob die postulierten Vorteile tatsächlich einen Effizienzgewinn erzielen, ist aufgrund der jungen Geschichte semantischer Technologien empirisch nur schwach belegt¹⁸⁴.

¹⁷⁹HAUSSER: *Grundlagen der Computerlinguistik: Mensch-Maschine-Kommunikation in natürlicher Sprache*, 2000, S. 410

¹⁸⁰Epimenides war ein Kreter, der den unsterblichen Satz aussprach: "Alle Kreter sind Lügner"

¹⁸¹SEBASTIANI: *A Note on Logic and Information Retrieval*, 1996, S. 13

¹⁸²STUDER/SCHNURR/NIERLICH: *Semantisches Knowledge Retrieval*, 2001, S. 14

¹⁸³vgl. JACKSON/MOULINIER: *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, 2002, S. 50-52

¹⁸⁴SCHMALTZ/HAGENHOFF: *Semantic Web Technologien für das Wissensmanagement*, 2004

2.6 Informationsvisualisierung im Information Retrieval

Die Informationsvisualisierung kann spätestens seit der Einführung graphischer Nutzerschnittstellen in Computersysteme als ein eigenständiges Teilgebiet des Forschungsbereichs Mensch-Technik Interaktion (MTI) betrachtet werden. Sie beschäftigt sich mit der bildlichen Darstellung von abstrakten, meist mehrdimensionalen Daten wie Dateisystemen oder Hypermediastrukturen, die keine direkten Entsprechungen in der "wirklichen" Welt besitzen. Daher besteht eine Abgrenzung zu der so genannten wissenschaftlichen Visualisierung, deren Gegenstandsbereich die bildliche Darstellung von physikalischen Daten wie Wetterphänomenen, Verbrennungsvorgänge oder Strömungen beinhaltet. Einen starken, nicht aus der Informatik stammenden Einfluss auf die Informationsvisualisierung hat das Gebiet des Designs. Vor allem Jacques Bertin und Edward Tufte sind hier als einflussreiche Personen zu nennen. Weitere Einflüsse kommen aus der Kognitionspsychologie, besonders aus dem Themenbereich der menschlichen Bildverarbeitung.

Die menschliche Wahrnehmung ist eng mit Grafiken und visueller Information verknüpft¹⁸⁵. Bilder können sehr ansprechend sein, besonders wenn sie gut gestaltet sind. So werden 80% der Sinneseindrücke durch das Auge erfasst. So können Visualisierungen im Allgemeinen die menschliche Kapazität zur Informationsverarbeitung erhöhen¹⁸⁶. Die biologisch beschränkte Verarbeitungskapazität kann einerseits durch abstrakte gedankliche Strukturen wie Metawörter, Metamodelle oder Metaprozesse, sowie durch externe Hilfsmittel erweitert werden. Visualisierungen sind ein solches externes Hilfsmittel der Kognition, die es erlaubt, die Fähigkeiten des menschlichen Gehirns zu erweitern. Visualisierungen dienen dabei, ähnlich Notizen oder Bücher, als externe Arbeitsspeicher, die das menschliche Arbeitsgedächtnis ergänzen und entlasten. Erreicht werden kann dies zum einen durch einen wesentlich höheren Informationsgehalt, der mit einem Blick erfassbar ist und zum anderen durch interaktive Operationen, mit denen diese Information direkt über ihre Darstellung bearbeitet werden kann.

Bei der Gestaltung einer Informationsvisualisierung sollen sechs Ziele erreicht werden¹⁸⁷: Der/dem BetrachterIn sollen mehr Speicher- und Verarbeitungskapazitäten zur Verfügung gestellt und die Suchzeiten bis zum Finden der gewünschten Information verringert werden. Durch die visuelle Wahrnehmung sollen Schlussfolgerungsprozesse zur Entdeckung von Zusammenhängen unterstützt werden, die das Treffen von Entscheidungen erleichtern. Weiterhin sollen bildliche Mechanismen die Überwachung von Veränderungen der zugrunde liegenden Information ermöglichen und dazu diese Information innerhalb eines manipulierbaren Mediums enkodieren.

Ein bekanntes Referenzmodell der Informationsvisualisierung¹⁸⁸, das in Abbildung 2.7 dargestellt ist, zeigt schematisch den Ablauf, wie Rohdaten bis zu der eigentlichen Visualisierung aufbereitet werden. Die Ausgangsdaten werden interpretiert und mittels einer geeigneten Trans-

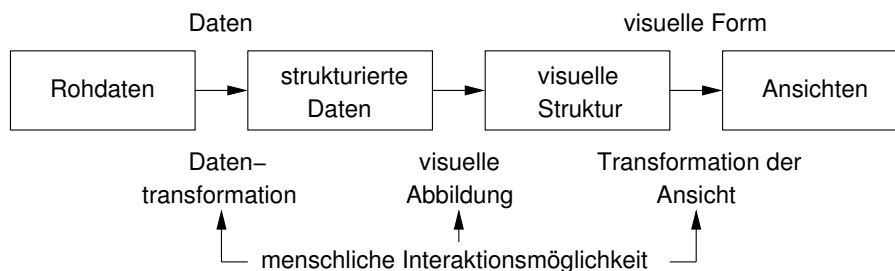


Abbildung 2.7: Referenzmodell der Informationsvisualisierung

formation in eine strukturierte Form wie relationale Tabellen oder Objektmodelle überführt. Diese Darstellung ist noch sehr datenzentriert. Die Datentabelle oder das Objektmodell stellt aber ei-

¹⁸⁵TUFTE: *The Visual Display of Quantitative Information*, 1983

¹⁸⁶vgl. CARD/MACKINLEY/SHNEIDERMAN: *Readings in Information Visualization: Using Vision to Think*, 1999, S. 10

¹⁸⁷vgl. CARD/MACKINLEY/SHNEIDERMAN: *Readings in Information Visualization: Using Vision to Think*, 1999, S. 16

¹⁸⁸vgl. CARD/MACKINLEY/SHNEIDERMAN: *Readings in Information Visualization: Using Vision to Think*, 1999, S. 17

ne spezielle, logische Sicht der Anwendung auf die Ausgangsdaten dar. Durch eine Abbildung werden die einzelnen Attribute der Daten in Visualisierungsstrukturen umgewandelt. Diese Abbildung stellt die Grundlage für die Darstellung der einzelnen Attribute dar. Sie können zum Beispiel als Koordinate, Farbe oder Muster repräsentiert werden. In einem weiteren Abbildungsschritt erfolgt die Umwandlung der Visualisierungsstruktur in eine oder mehrere Sichten. Dies kann durch einfache Transformationen wie affine Abbildungen oder Bildung von Ausschnitten erfolgen, aber auch durch Integration mehrerer Visualisierungstechniken.

Eine optimale Visualisierungstechnik für ein IR-System sollte alle notwendigen Interaktionen berücksichtigen und unterstützen. Dabei werden die folgenden Interaktionsphasen mit einem solchen System unterschieden¹⁸⁹:

1. Identifikation des Informationsbedürfnisses
2. Auswahl des Suchsystems und des Informationsraums
3. Formulierung der Anfrage
4. Anfrage beim System
5. Entgegennahme des Suchergebnisses in Form von Informationsobjekten
6. Inspektion, Evaluierung und Interpretation der Suchergebnisse
7. Eventuelle Reformulierung und Verfeinerung der Anfrage

Die durchzuführenden Suchaufgaben fallen in ein Spektrum zwischen der Beantwortung einer spezifischen Frage und der allgemeinen Recherche eines Themengebiets. Eine basale Unterstützung der erwähnten Interaktionsschritte findet sich in den meisten Internet-Suchmaschinen und in kommerziellen Systemen in Form von den drei Stufen: In einem Texteingabefeld können Suchanfragen in der jeweiligen Anfragesprache formuliert werden. Typischerweise ist dies eine boolesche Verkettung einzelner Suchterme. Durch Druck auf einen Anfrageknopf vollzieht das IR-System die Suche und gleicht die Anfrage mit dem Dokumentindex ab. Die Präsentation der Suchergebnisse erfolgt dann in einer nach vermuteter Relevanz sortierten Liste von Ergebnissen. Bei Hypertextsystemen beinhaltet ein Listeneintrag den Verweis auf das gefundene Dokument und oft noch eine kurze Darstellung des textuellen Kontexts in dem ein Suchbegriff auftaucht, wobei der Suchterm hervorgehoben wird.

Dem gegenüber stehen digitale Verzeichnisse, welche auf eine explizite Eingabe von Suchbegriffen verzichten. Dokumente oder Webseiten sind dort nach Themen hierarchisch klassifiziert. Solche Verzeichnisse wurden im Umfeld des weltweiten Netzes Anfang der 90er von Yahoo¹⁹⁰ händisch entwickelt. Weitergeführt wurde diese Arbeit von Altavista¹⁹¹, Google¹⁹² und neueren Datums von dem Open Directory Project¹⁹³. Das Open Directory Project nimmt dabei eine Sonderstellung ein, da sich dort jedermann bei der Klassifikation beteiligen kann. Ein solches Verzeichnis ist ähnlich einer Systematik aus dem Büchereiwesen durch die manuelle Selektion der Inhalte von hoher Qualität. Der Nutzer erhält damit eine navigationale Struktur, durch die er ausgehend von allgemeinen Themen bis zu einem gewünschten speziellen Thema vordringen und in diesem stöbern kann. Solche digitalen Verzeichnisse können als Themennetze beziehungsweise als nicht formalisierte Ontologien angesehen werden, da oft noch Querbeziehungen zwischen Themen hinterlegt sind.

¹⁸⁹BAEZA-YATES/RIBEIRO-NETO: *Modern Information Retrieval*, 1999, S. 263

¹⁹⁰<http://www.yahoo.de>

¹⁹¹<http://www.altavista.de>

¹⁹²<http://directory.google.de>

¹⁹³<http://www.dmoz.org>

2.6.1 Visualisierungen von vernetzten Begriffssystemen

Vernetzte Begriffssysteme, wie sie durch Ontologien formalisiert werden, stellen zentrale Datenstrukturen des Semantischen Netzes dar. Informationsnetze werden aber auch bei Anfragehilfen von Suchmaschinen und zur Kontextualisierung von Suchergebnissen eingesetzt. Hierbei stellt allerdings die Visualisierung und Navigation solcher Netzstrukturen ein erhebliches Problem bei der Gestaltung geeigneter Benutzungsschnittstellen dar. Gestaltungsfragen beziehen sich etwa auf die Minimierung der visuellen Suchpfade, die Übersichtlichkeit der Gesamtstruktur oder die Effizienz interaktiver Explorationsmöglichkeiten.

Eine gewöhnliche, graphähnliche Visualisierung wurde innerhalb des OntoViz-Moduls des Ontologie-Autorenwerkzeug Protege-2000¹⁹⁴ verwendet. Es basiert auf GraphViz¹⁹⁵, einer von AT&T entwickelten Umgebung zur Graphenanordnung und -visualisierung. Dieser Typ der visuellen Repräsentation besitzt aber den Nachteil, dass nur ein kleiner Teil des gesamten Graphen dargestellt werden kann, was die Übersichtlichkeit reduziert.

Für eine bessere Übersichtlichkeit wäre ein nahe liegender Ansatz, zwei oder mehrere Ansichten zu bieten, von denen eine den ganzen Graphen und die andere(n) vergrößerte Teilausschnitte zeigen. Dies hätte den Vorteil, dass sowohl lokale Details als auch die gesamte Struktur sichtbar wäre. Allerdings hätte dieses Vorgehen auch den Nachteil, mehr Platz auf dem Bildschirm zu erfordern und den Betrachter dazu zu nötigen, alle diese Ansichten mental zu integrieren. Naturgemäß sind auch die an die vergrößerten Teile des Graphs angrenzenden Gebiete nicht sichtbar, was ein Blättern des Bildes notwendig machen würde.

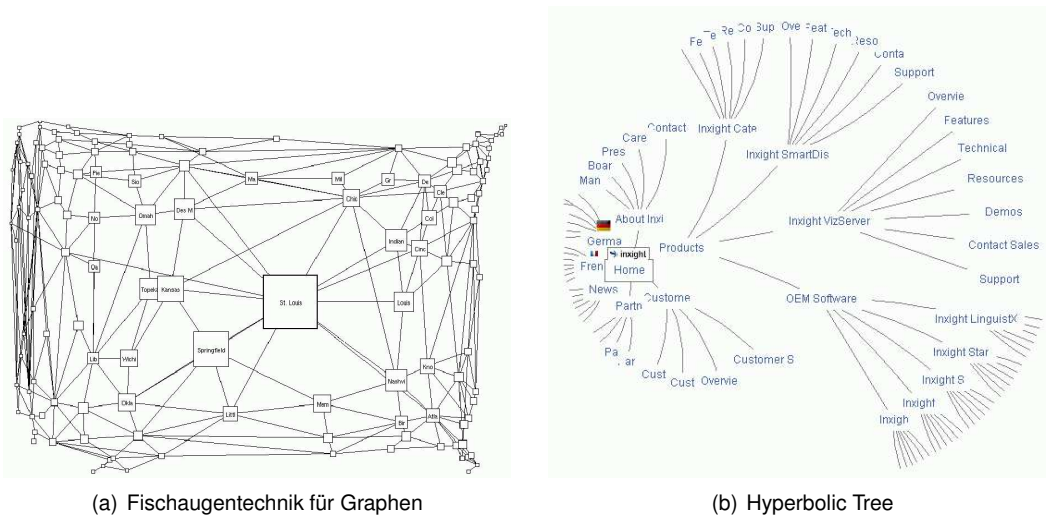


Abbildung 2.8: Verzerrungsbasierte Visualisierungstechniken mit Fischaugeneffekt

Mit der Intention Fokus und Kontext in einer Ansicht zu integrieren, wurden eine Reihe Visualisierungstechniken mit Fischaugeneffekt vorgestellt. Diese gehen alle auf das Pionierwerk von Furnas¹⁹⁶ zurück. Zum Beispiel bietet die Technik von Sakar¹⁹⁷ zur Darstellung und Anordnung von Graphen eine kontinuierliche und direkt-manipulative Interaktion (siehe Abbildung 2.8 (a)). Der Graph kann damit zwar reibungslos exploriert, jedoch hierarchische Beziehungstypen nicht sofort erkannt werden.

Mit einer anderen Fischaugentechnik für die Visualisierung großer Bäume, dem Hyperbolic-Tree¹⁹⁸ (siehe Abbildung 2.8 (b)), wurde die herkömmliche Art ersetzt, eine euklidische Fläche

¹⁹⁴ NOY et al.: *Creating Semantic Web Contents with Protege-2000*, 2001

¹⁹⁵ GANSNER/NORTH: *An Open Graph Visualization System and its Applications to Software Engineering*, 1999

¹⁹⁶ FURNAS: *Generalized Fisheye Views*, 1986

¹⁹⁷ SARKAR/BROWN: *Graphical Fisheye Views of Graphs*, 1992

¹⁹⁸ LAMPING/RAO: *Laying Out and Visualizing Large Trees Using a Hyperbolic Space*, 1994

zur Darstellung zu verwenden. Bei dieser Visualisierungstechnik werden erst die Knoten und hierarchische Kanten des Baumes auf eine hyperbolische Fläche platziert und diese dann auf die ebene Darstellungsfäche projiziert. Aufgrund des Verzerrungsmechanismus dieses Ansatzes sind jedoch nur Beziehungen zwischen räumlich benachbarten Knoten im Anzeigefokus vollständig zu erkennen. Obwohl er eigentlich nur für die Visualisierung hierarchischer Strukturen geeignet ist, wurde er jedoch in einigen Fällen auch für die Darstellung von Themenkarten und Ontologien eingesetzt. Ein Beispiel hierfür ist das später noch detaillierter vorgestellte semantische Suchsystem *Ontobroker*¹⁹⁹.

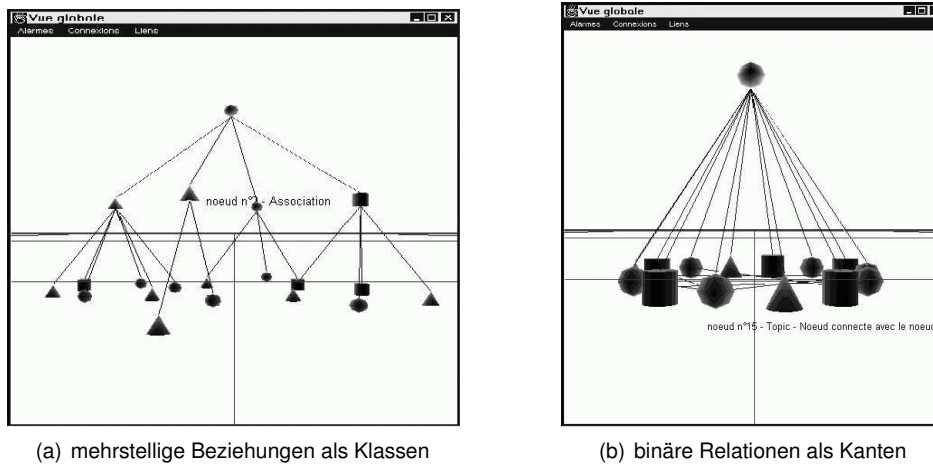


Abbildung 2.9: 3D-Visualisierung von Themenkarten

Ein dreidimensionales Visualisierungsverfahren²⁰⁰ für XML Topic Maps, welches auch zur Darstellung von Ontologien herangezogen werden könnte, zeigt Abbildung 2.9. Es beruht auf Kegelbäumen²⁰¹, die um interaktive Möglichkeiten erweitert wurden. Die Anzahl der dargestellten Knoten kann durch Filterungs- und Klassifizierungsalgorithmen reduziert werden. Mittels zweier Sichten kann die visualisierte Themenkarte erkundet werden. Zum einen können mehrstellige Beziehungen zwischen Knoten als Klassen dargestellt und zum anderen binäre Relationen als tatsächliche Kanten repräsentiert werden.

Alternativ wurden statische Netzvisualisierungen einschließlich der Repräsentation als Matrix schon von Bertin²⁰² vorgeschlagen. Interaktive Möglichkeiten wurden jedoch in dieser Arbeit nicht betrachtet.

2.6.2 Visualisierungen zur Unterstützung der Anfrageformulierung

Zur Unterstützung der Anfrageformulierung einer stichwortbasierten Suche kann die Visualisierung eines Thesaurus, wie von Fowler et al.²⁰³ vorgeschlagen, herangezogen werden. Aus einer natürlichsprachlichen Anfrage werden die Suchterme extrahiert, diese graphisch präsentiert und damit eine Stichwortsuche gestartet. Die Suchergebnisse können als Liste von textuellen Zusammenfassungen oder als Netzwerk von ähnlichen Dokumenten in Form eines Graphen dargestellt werden. Gleichzeitig werden durch einen zugrunde liegenden Thesaurus assoziierte Terme ebenfalls als Graph visualisiert. Diese assoziierten Terme können nun auf eine direkt-manipulative Art zu der Anfrage hinzugefügt oder von dieser entfernt werden. Weitere Hilfen finden sich auch in der Unterstützung zur logischen Verknüpfung von Suchbegriffen in booleschen Anfragen.

¹⁹⁹FENSEL et al.: *Ontobroker: The Very High Idea*, 1998

²⁰⁰LE GRAND/SOTO: *Information management - Topic maps visualization*, 2000

²⁰¹ROBERTSON/MACKINLAY/CARD.: *Cone trees: Animated 3D Visualizations of hierarchical Information*, 1991

²⁰²BERTIN: *Graphics and Graphic Information-Processing*, 1981

²⁰³FOWLER/FOWLER/WILSON: *Integrating Query Thesaurus and Documents through a visual Representation*, 1991

In dem System *VQuery*²⁰⁴ werden dazu zuerst eine Reihe von Suchtermen spezifiziert. Diese werden zusammen mit der Anzahl der jeweils damit erzielten Treffer mit Hilfe eines Venn-Diagramms visualisiert. Venn-Diagramme werden häufig in der Mengenlehre verwendet. Mengen werden dabei als Kreise und Schnittmengen als sich überlappende Kreise repräsentiert. Durch Verschieben der Kreise, welche die Suchterme repräsentieren, können diese zu einer Anfrage verknüpft werden. Überlappen sich Kreise, so handelt es sich dabei um mengenmäßige Konjunktion, während Kreise, die sich nicht überlappen, als ODER-Verknüpfung der korrespondierenden Terme angesehen werden. Auch können Kreise speziell markiert werden, um einen NICHT-Funktor auszudrücken.

Die Abfrage von Ontologien in wissensbasierten Systemen erfolgt prinzipiell in der Logiksprache, mit der sie verarbeitet werden. Dies ähnelt stark der strukturierten Abfrage von relationalen Datenbanksystemen. In Suchmaschinen im Umfeld des Semantischen Netzes wurden daher visuelle und interaktive Hilfen geschaffen, die den Nutzer bei der Formulierung einer Anfrage unterstützen. Dabei wird meist von der Annahme ausgegangen, der Nutzer interessiere sich für Instanzen von Klassen der Ontologie, die bestimmten Bedingungen genügen. In dem auf *SHOE* basierenden Suchwerkzeug²⁰⁵ wird hierzu zunächst eine Domänenontologie, welche als Suchkontext dient, ausgewählt. Die Klassen der Ontologie werden in Form einer hierarchischen, blätterbaren Liste visualisiert, wobei Subklassen eingerückt werden (siehe Abbildung 2.10 (a)). Nach der Auswahl einer Klasse werden deren Attribute und Relationen zu anderen Klassen in einer weiteren Liste angezeigt. Die Werte von Attributen, sowie die Namen von Klassen, die mit der selektierten in Beziehung stehen, kann nun sukzessive vom Nutzer eingegeben werden. Zur Erhöhung der Wahrscheinlichkeit von Suchtreffern wird in den Eingaben Groß- und Kleinschreibung nicht beachtet. Für eine Suchanfrage werden die eingegebenen Randbedingungen in einen logischen Ausdruck gebracht, der von zu findenden Instanzen erfüllt sein muss.

Ein ähnliches Interaktionsmuster verfolgt die Suchschnittstelle von *Ontobroker*²⁰⁶, einer Vermittlungsarchitektur für ontologische Information. Auch hier bildet der Suchkontext eine zuerst ausgewählte Ontologie. Diese kann mit Hilfe der in Abbildung 2.8 (b) gezeigten, hyperbolischen Darstellung des Ontologiegraphen exploriert werden. Zur Reduktion der dargestellten Information werden nur die Klassen der Ontologie und deren Attribute als Knoten des Graphen herangezogen. Die Kanten werden durch Subklassenrelationen und andere, typisierte Beziehungen zwischen Klassen gebildet. Zur eigentlichen Suche steht eine interaktive Tabelle zur Verfügung, deren Spalten gemäß dem Schema "Variable", "Klasse", "Relation" und "Wert" aufgebaut werden. In den Zeilen der Tabelle werden die entsprechenden Werte eingetragen. Aus jeder Zeile werden logische Ausdrücke generiert, denen die Instanzen der gewählten Klassen mit deren Randbedingungen genügen müssen. Die Ergebnisse jeder Zeile werden in den entsprechenden Variablen festgehalten, die wiederum mit booleschen Junktoren verknüpft werden können. Wird eine Klasse in der Graphenvisualisierung ausgewählt, so wird sie in die gerade aktive Zeile der Tabelle übernommen. Mögliche Beziehungen und Attribute der gewählten Klasse werden aus der Ontologie übernommen und können durch Auswahl aus einem Listenfeld mit restriktierenden Werten versehen werden. Handelt es sich bei der Relation um eine Beziehung zu einer anderen Klasse, kann diese Klasse, im Unterschied zu *SHOE*, im Graphen selektiert werden. Da mittels Relationen auch Attribute dargestellt werden, muss in diesem Fall der zugehörige Wert jedoch wiederum textuell eingegeben werden. Zur Kontrolle der Anfrage wird diese in der nativen Abfragesprache F-Logik zusätzlich präsentiert.

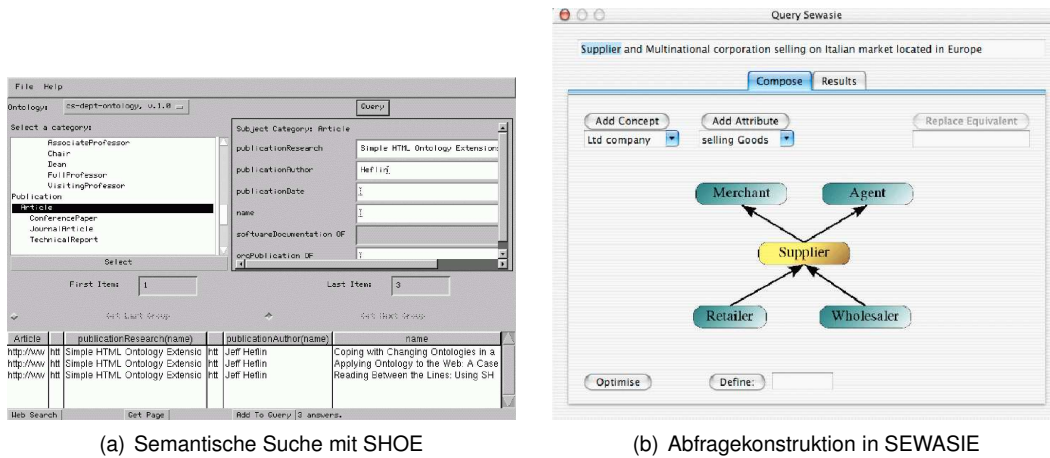
Ein einfacherer Ansatz zur Formulierung von Abfragen wurde in dem SEWASIE (SEmantic-Webs and AgentS in Integrated Economies) Projekt entwickelt²⁰⁷. Der Suchkontext ist hierbei implizit durch eine festgelegte, vom Anwendungsfall abhängige Domänenontologie gegeben. Als Ausgangspunkt zur Konstruktion einer Abfrage dienen eine Menge vordefinierter Suchanfragen, die weiter spezialisiert werden. Prototypische Anfragen haben den Vorteil, dass dem Nutzer nicht die Ontologie in ihrer Gesamtheit, sondern nur der für die Anfrage relevante Teil präsentiert werden muss. Damit entfällt der Navigationsaufwand zur Suche gewünschter Klassen. Die Klasse,

²⁰⁴ JONES: *Graphical Query Specification and dynamic Result Previews for a Digital Library*, 1998

²⁰⁵ HEFLIN/HENDLER: *Searching the Web with SHOE*, 2000

²⁰⁶ FENSEL et al.: *Ontobroker: The Very High Idea*, 1998

²⁰⁷ CATARCI et al.: *An Ontology Based Visual Tool for Query Formulation Support*, 2003



(a) Semantische Suche mit SHOE

(b) Abfragekonstruktion in SEWASIE

Abbildung 2.10: Formulierung semantischer Suchanfragen

nach deren Instanzen gesucht wird, stellt den Suchfokus dar. Deren Sub- und Superklassen werden wie in Abbildung 2.10 (b) in Form eines graphischen Baumes visualisiert. Durch Selektion einer dieser Sub- oder Superklassen kann der Suchfokus verschoben werden. Zusätzliche Randbedingungen, welche die Instanzen der Klasse erfüllen müssen, werden wie in *Ontobroker* aus einer Liste möglicher Relationen und Attributen ausgewählt. Die Instanzen können in dieser Schnittstelle noch weiter restriktiert werden, indem sie zusätzlich noch Instanzen weiterer Klassen sein müssen. Aus einer weiteren Liste können diese zusätzlichen Klassen ausgewählt werden.

2.6.3 Visualisierungen zur Kontextualisierung von Suchergebnissen

Bei der Interaktion mit Suchsystemen ist dem Nutzer oft nicht klar, wie und warum die gezeigten Suchergebnisse mit der Anfrage übereinstimmen und auf welcher Basis eine Sortierung erfolgt. Auch wird meist der Zusammenhang der Treffer untereinander nicht deutlich. Eine Einordnung in einen Kontext, der den thematischen Zusammenhang der gefundenen Dokumente verdeutlicht, gibt Hinweise sowohl auf deren Relevanz im Hinblick auf das Informationsbedürfnis als auch darauf, wie Suchterme bei einer erneuten Anfrage besser gewählt werden können.

In *SQWID*²⁰⁸ wird auf Basis einer stichwortbasierten Internetsuchanfrage eine dynamische Visualisierung der daraus resultierenden Webseiten generiert (vgl. Abbildung 2.11 (a)). Gleichzeitig werden eine Reihe anderer Terme, welche in Bezug zu der Anfrage stehen, aus den Webseiten der Ergebnismenge bestimmt. Aus diesen Termen und den gefundenen Seiten wird eine Graphenvisualisierung generiert. Die drei wichtigsten, durch ein Termgewichtungsverfahren bestimmten, Begriffe werden als Knoten des Graphen in einem Dreieck gruppiert. Die Suchergebnisse werden ebenfalls als Knoten dargestellt und in einer gewissen Distanz zu den Begriffen platziert. Die Distanz errechnet sich aus der Relevanz des Begriffes zu seinem korrespondierenden Ergebnis. Je höher dabei der Relevanzwert zu dem jeweiligen Begriff ist, desto näher werden sie zu diesem platziert. Daher hängt die Position eines Suchergebnisses von der Relevanz zu allen drei Begriffen ab. Wenn keine Zuordnung eines Ergebnisses zu den Begriffen erkannt werden konnte, werden sie am Rande der Ansicht positioniert. Die Kanten zwischen den Suchergebnisknoten und den Begriffsknoten werden mit dem jeweiligen Relevanzwert beschriftet. Durch einen Schieberegler kann die Anzahl der dargestellten Ergebnisse variiert werden.

Das in Abbildung 2.11 (b) gezeigte Suchsystem *kartoo* (<http://www.kartoo.com>) vernetzt ähnlich wie *SQWID* gefundene Webseiten. Jedoch werden diese mittels der extrahierten Begriffe untereinander in Beziehung gesetzt. Ein oder mehrere Stichwörter verbinden dabei Dokumente,

²⁰⁸ MCCRICKARD/KEHOE: *Visualizing Search Results using SQWID*, 1997

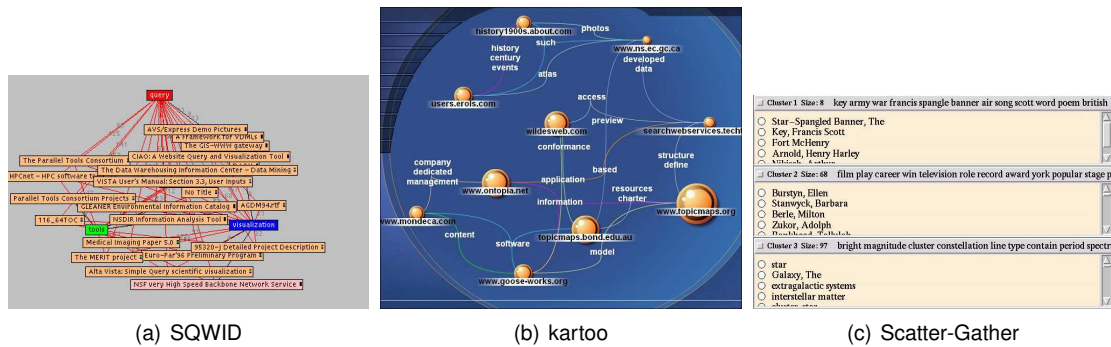


Abbildung 2.11: Visualisierungen zur Erklärung der Ergebnismenge

für welche die Stichwörter als relevant eingestuft wurden. Ein weiterer Unterschied zu SQWID besteht darin, dass die verbindenden Terme zur nutzergesteuerten Anfrageverfeinerung genutzt werden können. Durch interaktive "+" und "-" Symbole können die gezeigten Terme durch die booleschen Junktoren "UND" und "NICHT" mit der Anfrage verknüpft werden.

Ein ähnliches Ziel, jedoch mit einer anderen Interaktionsform, verfolgt die in Abbildung 2.11 (c) gezeigte Schnittstelle *Scatter-Gather*²⁰⁹ zum Stöbern in umfangreichen Dokumentsammlungen. Die Ergebnisdokumente einer Stichwortwortsuche werden aufgrund ihrer Indexterme zu einer nutzerdefinierten Anzahl von thematischen Gruppen zusammengefasst. Jede dieser Gruppen wird mit einer Reihe von für diese Gruppe wichtigen Begriffen beschrieben und zusammen mit der Anzahl der Dokumente in einem blätterbaren Teilausschnitt der Ansicht visualisiert. Der Nutzer erhält so eine Möglichkeit den ungefähren thematischen Inhalt der in einer Gruppe zusammen gefassten Dokumente zu errahnen. Entweder kann daraus ein Dokument ausgewählt werden oder der Gruppierungsvorgang auf Basis der Untermenge der in dieser Gruppe enthaltenen Dokumente neu gestartet werden.

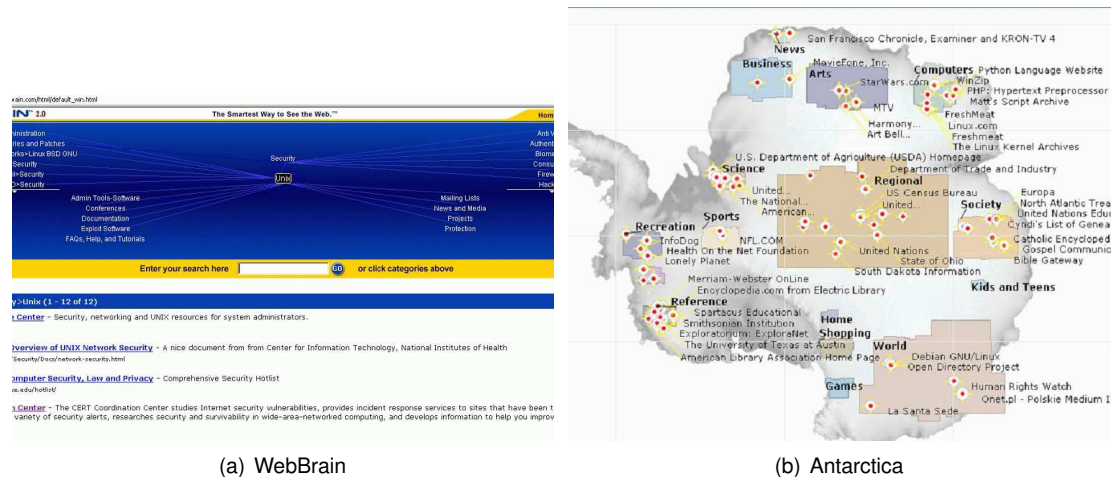


Abbildung 2.12: Visualisierungen zur Kontextualisierung von Suchergebnissen

Mit dem Vorhandensein elektronischer Verzeichnisse konzentrieren sich neuere Visualisierungsarten wie *The Brain* (<http://www.thebrain.com>) oder *Antarctica*²¹⁰ (<http://maps.map.net>) auf eine Kontextualisierung der Suchergebnisse in vordefinierten Themengebieten. Gefundene Dokumente werden thematischen Kategorien zugeordnet und interaktiv präsentiert. Wieder mit Hilfe

²⁰⁹ CUTTING et al.: *Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections*, 1992

²¹⁰ beide PATIENCE/CHALMERS: *Unstructured Data Management: the Elephant in the Corner*, 2002

einer Graphendarstellung visualisiert *The Brain* eine gegebene Themenhierarchie. Jedes Thema stellt dabei, wie in Abbildung 2.12 (a), einen Knoten im Graph dar. Die Kanten des Graphen repräsentieren einen hierarchischen Beziehungstyp von Oberthema zu Unterthema. Um einen Zentrums-knoten herum, welcher ein Oberthema verkörpert, werden alle Subthemen ringförmig angeordnet. In einem separaten, blätterbaren Bereich werden die zu dem Oberthema zugehörigen Dokumente gezeigt. Durch die Selektion eines Subthemenknoten wandert dieser in die Mitte und stellt den neuen Oberthemenknoten dar. Dessen Subthemen werden wiederum kreisförmig darum herum angeordnet und die zugehörigen Dokumente dargestellt. Auf diese Weise kann themenbezogen im Dokumentenraum gestöbert werden. Neben diesem Stöbermechanismus besitzt *The Brain* auch Suchfunktionalitäten. Wird mit Stichwörtern gesucht, so zeigt *The Brain* dasjenige Themengebiet als Zentrums-knoten an, in dem die meisten Treffer erzielt wurden. Zusätzlich werden alle Dokumente diesen Themas angezeigt, ohne jedoch die Suchtreffer speziell zu kennzeichnen.

Das System *Antarctica* verwendet die Metapher eine geographischen Landkarte, die den Zusammenhang zwischen Themengebieten und Dokumenten darstellt. Auch diese Art der Visualisierung erlaubt sowohl ein Stöbern als auch eine Suche anhand von Stichwörtern. In *Antarctica* werden die verschiedenen Themengebiete als Länder auf einem virtuellen Kontinent dargestellt, wie das Bildschirmfoto in Abbildung 2.12 (b) zeigt. In den Ländern sind die Dokumente als Städte enthalten. Wird ein bestimmtes Thema angewählt, so reagiert die Landkarte mit einem Zoomvorgang in dieses Gebiet. Ähnliches Verhalten zeigt eine Stichwortsuche. Dabei wird die Region angezeigt, in der die meisten Treffer erzielt wurden. Die Kartenmetapher wurde schon in anderen IR-Systemen wie zum Beispiel *WebSOM*²¹¹ vorgeschlagen. Jedoch verwenden diese Systeme kein explizites Verzeichnis, sondern stützen sich ähnlich wie *Scatter-Gather* auf Textgruppierungsverfahren.

2.6.4 Diskussion

Das herkömmliche Interaktionsmuster von Suchmaschinen besteht in der Eingabe einer Suchanfrage in Form einer Verkettung von Stichwörtern, der eigentlichen Suchanfrage und der Präsentation der Suchergebnisse, meist dargestellt als flache Tabelle. Eine Zeile dieser Tabelle enthält neben der Referenz auf das jeweilige Dokument zumindest dessen Titel, manchmal aber auch Textpassagen, in denen die Suchwörter auftauchen. Anhand dieser Information muss diese Tabelle vom Nutzer auf Hinweise nach relevanten Dokumenten durchsucht werden. Meist müssen die referenzierten Dokumente jedoch direkt inspiziert werden, um die Relevanz abschätzen zu können. Entsprechen die Suchergebnisse nicht dem eigentlich Gesuchten, so erfolgt ein Iterationszyklus, in dem die Anfrage reformuliert und erneut beim System angefragt wird. Zudem erfolgt damit ein dreimaliger Kontextwechsel, der seitens des Nutzers mental integriert werden muss.

Die tabellenbasierte Darstellung der gefundenen Dokumente bietet wenig Unterstützung beim Durchsuchen und der Erkennung von Zusammenhängen in der Ergebnismenge. Die Inspektion der Tabelle muss linear von Zeile zu Zeile erfolgen, welche aufgrund eines spezifischen Algorithmus zur Bestimmung der Rangfolge der jeweiligen Suchmaschine vorsortiert wurde. Das Durchsuchen der Tabelle wird zusätzlich behindert, wenn nur ein kleiner Ausschnitt der Tabelle (ca. 20 Einträge) dargestellt wird, wie es im Umfeld des WWW üblich ist. Daher muss weiter geblättert werden, wobei normalerweise auf die Systemantwort gewartet werden muss. Bei umfangreichen Informationsräumen ist es mit dieser Art der Interaktion fast unmöglich ein Verständnis von dessen Struktur und Inhalt zu entwickeln, da die Iterationszyklen sehr oft durchlaufen werden müssen, um Querbezüge zwischen einzelnen Inhalten herzustellen. Zudem erlaubt der rein stichwortbasierte Zugang nur die Erfüllung eines schmalen Spektrums von Suchaufgaben. Ganze Themengebiete oder spezifische Fragen lassen sich nur schwer recherchieren und bedingen ebenfalls etliche Iterationszyklen.

Eine thematische Recherche wird mit elektronischen Verzeichnissen möglich. Deren taxonomische Natur wird zur Präsentation jedoch meist auf eine Liste von Themengebieten einer Hierarchieebene reduziert. Zur Spezialisierung von Themen wird das Gewünschte ausgewählt.

²¹¹LAGUS et al.: *WebSOM for Textual Data Mining*, 1999

Das System präsentiert daraufhin die direkten Unterthemen des Ausgewählten. Daher gilt ähnliches dem oben ausgesagtem auch für das Stöbern in solcherart visualisierten Verzeichnissen. Zum Beispiel benötigt im Verzeichnis von *Google* das Vordringen von oberster Ebene zum Thema "Google" sechs Interaktionsschritte. In jedem dieser Schritte muss ein Unterthema aus einer Liste von circa 30 Einträgen ausgewählt werden. Dieses schrittweise Vorgehen behindert enorm die Erarbeitung einer Übersicht der gesamten Themenstruktur.

Diskussion von Visualisierungen vernetzter Begriffssysteme

Sowohl bei der visuellen Hilfe zur Anfrageformulierung als auch bei Darstellung von Zusammenhängen und Kontexten von Suchergebnissen werden oft netzartige Datenstrukturen herangezogen. Auch im Umfeld der semantischen Suche sind Begriffs- und Informationsnetze in Form von Ontologien gegeben. Die gebräuchlichste visuelle Struktur für Netze ist dabei die Repräsentation als Graph mit Knoten und Kanten. Diese Repräsentation wird zwar häufig verwendet, besitzt jedoch folgende Nachteile: Ist das Netzwerk groß und besitzt einen hohen Grad an Vernetzung, so wird es schwierig, den Überblick zu behalten und bestimmte Ausschnitte zu finden. Auch wird die visuelle Suche durch meist beliebige und unstrukturierte Anordnungen der Knoten behindert. Aus Platzgründen können oft auch assoziierte Knoten nur weit voneinander entfernt platziert werden.

Während es bei dieser Darstellungsform meist noch einfach ist, die direkte Nachbarschaft von Knoten zu erfassen, wird es schwierig und fehleranfällig, größere Beziehungsmuster zu erkennen oder visuell längeren Pfaden zu folgen. Keiner der vorgestellten Ansätze zur Graphvisualisierung unterstützt in ausreichender Form die Darstellung beliebig vernetzter Informationsstrukturen im Hinblick auf unterschiedliche Nutzeraufgaben wie die Suche spezifischer Knoten und Kanten, das Entdecken der Beziehungen zwischen beliebigen Knoten oder die Erfassung aller Beziehungen eines bestimmten Knotens. Weiterhin fehlen wesentliche interaktive Eigenschaften wie ein systematisches Vordringen in und die Verdichtung von Teilstrukturen. Aus diesen Gründen sind die existierenden Techniken nicht ausreichend geeignet, um komplexe Netzstrukturen zu visualisieren und zu explorieren²¹².

Diskussion von Visualisierungen zur Unterstützung der Anfrageformulierung

Die stichwortbasierte Suche erzeugt oft entweder zu viele Treffer, wenn das Informationsbedürfnis mit zu wenigen oder zu allgemeinen Begriffen ausgedrückt wird, oder zu wenig Treffer, wenn zu viele oder zu spezifische Terme zur Suche verwendet werden. Die Anfragekomposition mittels Venn-Diagrammen, in denen Suchbegriffe mit der zugehörigen Anzahl der Treffer gezeigt werden und die eine interaktive Verknüpfung der Anfrageterme erlauben, erzeugt schnellere und präzisere Suchergebnisse²¹³. Eine Hilfestellung bei der eigentlichen Wahl der Suchterme gibt die interaktive Darstellung eines Thesaurus. Der Nutzer erhält damit die Kontrolle über das auszuwählende Vokabular und die Umformulierung der Anfrage. Damit kann auch den Nachteilen der automatischen Anfrageerweiterung durch kontextirrelevante Terme begegnet werden. Ergonomische Probleme ergeben sich jedoch, wie oben gezeigt, bei der Visualisierung des Thesaurus als Graphen.

Eine besondere Bedeutung kommt der Hilfestellung bei der Konstruktion von logischen Abfragen zur Suche in Ontologien zu. Die Anfrage in der jeweiligen Repräsentationssprache der Ontologie erfordert sowohl das Erlernen und eine genaue Kenntnis dieser Sprache als auch ein vorheriges Verständnis des Inhalts und der Bedeutung des ontologischen Modells. Zudem muss im Vorfeld bekannt sein, wie die Klassen, Relationen und Attribute benannt wurden. Die Suchmaschine *SHOE* behilft sich mit einer Liste aller Klassennamen, bei der Subklassen zur Darstellung einer Taxonomie eingerückt werden. Eine Ontologie enthält jedoch auch Beziehungen zwischen Klassen, die wichtig für einen Überblick wären. Diese können in einer Liste nicht dargestellt werden. Mit zunehmender Anzahl der Klassen in einer Ontologie steigt auch der Interaktionsaufwand

²¹²vgl. ZIEGLER/KUNZ/BOTSCH: *Matrix Browser: Visualisierung und Exploration vernetzter Informationsräume*, 2002

²¹³vgl. BAEZA-YATES/RIBEIRO-NETO: *Modern Information Retrieval*, 1999, S. 283

zum Blättern in der Liste, deren Sortierung vorgegeben ist und die nicht gefiltert werden kann. Eine Eingrenzung von Instanzen einer gewählten Klasse erfolgt mittels der Spezifikation von Beziehungen zu Instanzen einer anderen Klasse beziehungsweise mit der Angabe von Attributen, die diese besitzen müssen. Eine graphische Hilfestellung zur Auswahl der Zielklasse wird in *SHOE* nicht geboten. Die Liste der Klassen muss weiter durchsucht und der Name der Zielklasse im Suchdialog textuell eingetragen werden. Auch wird im Suchdialog nicht gekennzeichnet, bei welchem Eingabefeld es sich um den Wert eines Attributs oder um die Relation zu einer anderen Klasse handelt.

Für eine bessere Übersichtlichkeit der visualisierten Ontologie wird in *Ontobroker* der HyperbolicTree verwendet. Für diesen Anwendungsfall ist er allerdings nur bedingt einsetzbar, da bei nichthierarchischer Vernetzungsstruktur manche Knoten mehrfach in der Darstellung auftauchen können. Zusätzlich existieren auch kontroverse Performanz- und Benutzbarkeitsbefunde²¹⁴²¹⁵. Die Präsentation der Ontologie und der Suchdialog sind im Gegensatz zu *SHOE* derart gekoppelt, dass sowohl die Abfrageklasse als auch die Zielklasse daraus übernommen werden kann. Lediglich der Wert eines Attributs muss eingegeben werden. Auch hier erfolgt bei der Auswahl der zu spezifizierenden Relationen und Attribute keine Kennzeichnung des jeweiligen Typs.

Zur Reduzierung der visualisierten Information bedient sich die Nutzerschnittstelle in *SEWA-SIE* vorgefertigter Suchanfragen, die jeweils nur einen Teilausschnitt mit wenigen Knoten der Ontologie zur Darstellung benötigen. Damit kann jedoch nur schwer ein Überblick über die gesamte Struktur der Ontologie gegeben werden, da der Suchfokus dazu immer weiter verschoben werden muss. Auch hier werden in der Visualisierung des Ontologiegraphen nur hierarchische Beziehungstypen berücksichtigt. Für die Auswahl von Relationen und Attributen gilt daher ähnliches wie bei *Ontobroker*. Vorgefertigte Suchanfragen besitzen nicht nur den Nachteil eines erhöhten Aufwandes bei deren Auswahl und Redaktion, sie sind auch unflexibel im Falle einer Übertragung auf andere Anwendungsfälle oder der Berücksichtigung nicht bedachter Informationsbedürfnisse.

Diskussion von Kontextualisierungsfunktionen

Einige Visualisierungstechniken wurden vorgestellt, die einen Einblick in den thematischen Kontext der Suchergebnisse bieten. Dabei lassen sich zwei grundsätzliche Ansätze unterscheiden: Systeme wie *SQWID*, *kartoo* und *Scatter-Gather* nutzen Termgewichtungsverfahren, um möglichst inhaltsrelevante Stichwörter zu extrahieren, während *The Brain* und *Antarctica* manuell erstellte Klassifikationen als thematischen Kontext nutzen. Extrahierte Stichwörter liefern jedoch nur Hinweise auf etwaige, allgemeine Themengebiete, da Themenbegriffe oft nicht direkt in den Texten verwendet werden. Aus einer Menge von Stichwörtern muss daher auf das eigentliche Thema geschlossen werden, was eine kognitive Mehrbelastung des Nutzers bedeutet.

SQWID nutzt wiederum eine Graphendarstellung für die Visualisierung der Zusammenhänge. Zwar können thematisch zusammenhängende Ergebnisdokumente visuell erkannt werden, aus Platzgründen kann jedoch nur ein geringer Teil der Treffer in der Darstellung berücksichtigt werden. Auch ist es kaum wahrscheinlich, dass drei Stichwörter genügen, um die Thematiken zu unterscheiden, die von einer großen Treffermenge behandelt werden. Eine höhere Skalierbarkeit erreicht *Scatter-Gather*, da hierbei hierarchische Themengruppen innerhalb der Ergebnismenge gebildet werden, die jeweils durch eine Menge von Stichwörtern beschrieben werden. Für jede Themengruppe muss aus diesen auf das allgemeine Thema dieser Gruppe geschlossen werden. Da jeweils nur eine Hierarchieebene gezeigt wird, kann nur durch hohen Navigationsaufwand ein Überblick über gefundene Themen erlangt werden. *kartoo* zeigt gefundene Dokumente in einem Netzwerk, wobei die Treffer durch extrahierte Stichwörter verbunden sind. Das System gibt jedoch keine Hinweise, wie die Ergebnismenge im Zusammenhang mit der Anfrage stand oder welche unterschiedlichen Thematiken behandelt werden. Aus Platzgründen können nur wenig Treffer berücksichtigt werden. Allen drei Systemen ist gemein, dass nur die Titel der angezeigten Dokumente dargestellt werden. Weitere Hinweise auf den Inhalt eines jeden Dokuments wird nicht gegeben.

²¹⁴PIROLLI/CARD/WEGE: *The Effect of Information Scent on Searching Information Visualizations*, 2000

²¹⁵NOLLER: *Ein Usability-Experiment zur Informationssuche im World-Wide-Web*, 2000

Die Verwendung von Verzeichnissen in *The Brain* und *Antarctica* hat den Vorteil, dass Themen direkt gegeben sind und nicht mit Stichwörtern beschrieben werden müssen. In *The Brain* wird auf eine Anfrage das Themengebiet bestimmt, in welchem die meisten Treffer erzielt wurden und alle Dokumente dieses Themas zurück geliefert. Daher wird die Anfrage nicht mit relevanten Dokumenten abgeglichen, sondern mit einem relevanten Thema. Dies hat eine geringere Vollständigkeit zur Folge, da möglicherweise nicht alle relevanten Dokumente gefunden wurden. Gleichfalls werden möglicherweise unwichtige Dokumente angezeigt, was den Relevanzwert senkt. Das System *Antarctica* visualisiert daher alle Themengebiete in denen Suchtreffer erzielt wurden. Jedoch können auch hier aus Platzgründen nur die vermeintlich wichtigsten Treffer berücksichtigt werden, von denen jeweils nur der Titel angezeigt wird.

Keine der vorgestellten Visualisierungsansätze unterstützt alle Interaktionsschritte, die in einer Recheresituation durchlaufen werden. Sie stellen vielmehr Insellösungen für einzelne Problemfälle dar. Weiterhin werden hauptsächlich graphenbasierte Visualisierungstechniken zur Darstellung von Zusammenhängen verwendet. Dies trägt zwar der vernetzten Informationsstruktur Rechnung, birgt aber die geschilderten Problempunkte.

3 Ein integrierter Ansatz zur wissensbasierten Informationsrecherche

Das Hauptziel der vorliegenden Arbeit ist die Entwicklung und Evaluation eines nutzerzentrierten und ganzheitlichen Ansatzes, welcher es erlaubt, große und vernetzte Informationsräume unter Einbeziehung der menschlichen Nutzung und Interpretation semantisch erschließ- und recherchierbar zu machen. Unter dem Begriff Informationsraum werden außer Dokumentkollektionen auch sonstige Ressourcen wie Dienste und Produktdaten zusammengefasst, deren Inhalt von Interesse ist. Ausgangspunkt ist die Annahme, dass in Situationen der Informationssuche, ein Kommunikationsprozess zwischen dem menschlichen Nutzer und dem benutzten Werkzeug statt findet, welcher es erlaubt aus der Vielzahl vorhandener Information die Passende zu selektieren und zu übertragen. Erst diese Annahme ermöglicht die Erarbeitung eines integrierten Ansatzes zur wissensbasierten Informationsrecherche auf den Ebenen der Gestaltung der inneren Beschaffenheit des Werkzeuges, das Aussehen der Nutzerschnittstelle und der Extraktion von Information aus diesem Kommunikationsprozess. Die extrahierte Information soll Lernprozesse und damit die Bildung von neuem Wissen für die beteiligten Kommunikationspartner ermöglichen. Den Anwendungskontext stellen Intranets, komplexe Portalseiten, sowie so genannte Wissensdatenbanken dar. Wissensdatenbanken sind spezialisierte Dokumentkollektionen, welche von Expertengruppen zur Dokumentation gemeinsamer Erfahrungen und bekannten Problemlösungen erstellt werden.

3.1 Anforderungen an einen integrierten Ansatz zur Informationsrecherche

Für die Entwicklung eines geeigneten Ansatzes zur Informationsrecherche ist es essentiell, das Problem des Nutzers bei dem Umgang mit einem Suchsystem zu verstehen. Dieses besteht vorrangig darin, den Informationsbedarf des Nutzers mit dem Informationsangebot des Systems zu stillen. Für das System bedeutet dies, den Bedarf des Nutzers zu erkennen und möglichst effizient auszugleichen. Im Anwendungsfall sollte auf eine Suchanfrage, durch die der Informationsbedarf ausgedrückt wird, jede relevante und keine überflüssige Information aufgefunden und als Suchergebnis angeboten werden. Überflüssige Information ist hier diejenige, welche nicht zum Ausgleich des Informationsbedarfs benötigt wird.

Der *Informationsbedarf* (vgl. Abbildung 3.1) des Suchenden ist definiert als "die Art Menge und Qualität der Information, die eine Person zur Erfüllung ihrer Aufgabe in einer bestimmten Zeit benötigt"¹. Er gliedert sich in einen objektiven und subjektiven Teil. Der so genannte *objektive Informationsbedarf* gibt an, welche Art und Menge von Information tatsächlich zur Erfüllung der Aufgabe benötigt wird. Der *subjektive Informationsbedarf* geht hingegen von der Sichtweise der jeweiligen Person aus und gibt diejenige Information an, welche dieser zur Bearbeitung der Aufgabe als relevant erscheint. Der objektive und subjektive Informationsbedarf ist oft stark unterschiedlich, vor allem bei schwer zu spezifizierenden, komplexen Aufgaben oder bei unbekanntem Strukturen und Zusammenhängen des Informationsraums. Die Menge an Information, die letztlich tatsächlich nachgefragt wird, ist wiederum lediglich eine Teilmenge des ursprünglich vermuteten Informationsbedarfs.

Nach dieser erweiterten Problematik ist es zudem Aufgabe des Systems, den subjektiven Informationsbedarf dem objektiven anzunähern. Nur der Bereich, in dem die tatsächliche Informationsnachfrage und das Informationsangebot zusammenfallen, führt schließlich zu einer konkreten

¹PICOT/REICHWALD/WIGAND: *Die Grenzenlose Unternehmung*, 2003, S.106

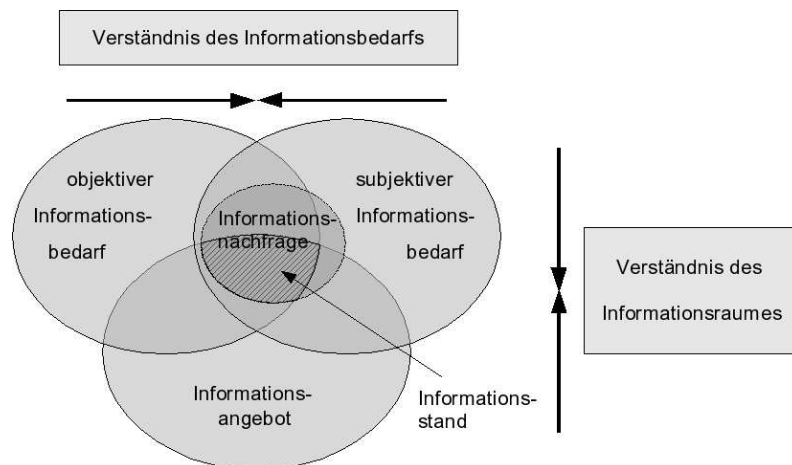


Abbildung 3.1: Das Suchproblem

Informationsversorgung. Der Teil der Informationsversorgung, der objektiv zur Aufgabenerfüllung notwendig ist, stellt den *Informationsstand* dar.

An einen Lösungsansatz sind daher die folgenden Anforderungen zu stellen:

1. Bereitstellung von effektiven Zugangs- beziehungsweise Abfragemechanismen, die alle Inhalte des Angebots schnell und präzise erreichen.
2. Förderung des Verständnisses von Informationsraum und eigenem Informationsbedarfs.
3. Hilfestellung bei der inhaltlichen Erfassung der Ergebnismenge einer Suchanfrage.

Effektive Zugangs- und Abfragemechanismen müssen angeboten werden, welche das gesamte Spektrum von Suchaufgaben unterstützen. Dieses reicht von der allgemeinen Recherche von Themengebieten bis zur gezielten Lokalisierung gewünschter Information (vgl. Abschnitt 2.6, S. 57). Ein möglichst gutes Verständnis des Informationsraums und des eigenen Informationsbedarfes ist vorteilhaft. Erst durch ein inhaltliches Begreifen des Informationsraums kann es gelingen, Anfragen so zu stellen, dass gewünschte Ergebnisse berechnet werden können. Jedoch die richtigen Anfragen zu stellen, welche zu einer tatsächlichen Informationsversorgung führen, wird nur durch ein Verstehen des eigenen objektiven Informationsbedarfes ermöglicht. Der objektive Informationsbedarf wiederum ist nur im Dialog und durch Rückkopplung unter den Kommunikationspartner zu ermitteln. Daher ist auch die Ergebnismenge einer Suchanfrage so aufzubereiten, dass sie leicht und schnell erfasst, sowie in ihrem Inhalt begriffen werden kann. Die Bandbreite der Kommunikation wird dadurch erhöht.

3.2 Grundlegender Ansatz und Begründung

Der gewählte Ansatz ist in Abbildung 3.2 als Übersicht dargestellt. Er integriert eine herkömmliche Stichwortsuchmaschine mit einer neuartigen Schnittstelle zwischen Mensch und Maschine, einer ebenfalls neuartigen Komponente zur Beobachtung des Ablageverhaltens der Nutzer, sowie einer probabilistischen Deduktionskomponente zu einem ganzheitlichen Rahmenwerk.

Das Hauptmerkmal der Nutzerschnittstelle ist die interaktive Visualisierung der semantischen Struktur des Informationsraums zur Unterstützung eines besseren Verständnisses desselben. Im Minimalfall beinhaltet die semantische Struktur ein dem Informationsraum zugehöriges Themen-Netzwerk, sowie bedeutungsvolle Merkmale von Dokumenten (Autor, Publikationsdatum, etc). Die Strukturvisualisierung ist mit einer Präsentation von Suchergebnissen als Tabelle zu einem

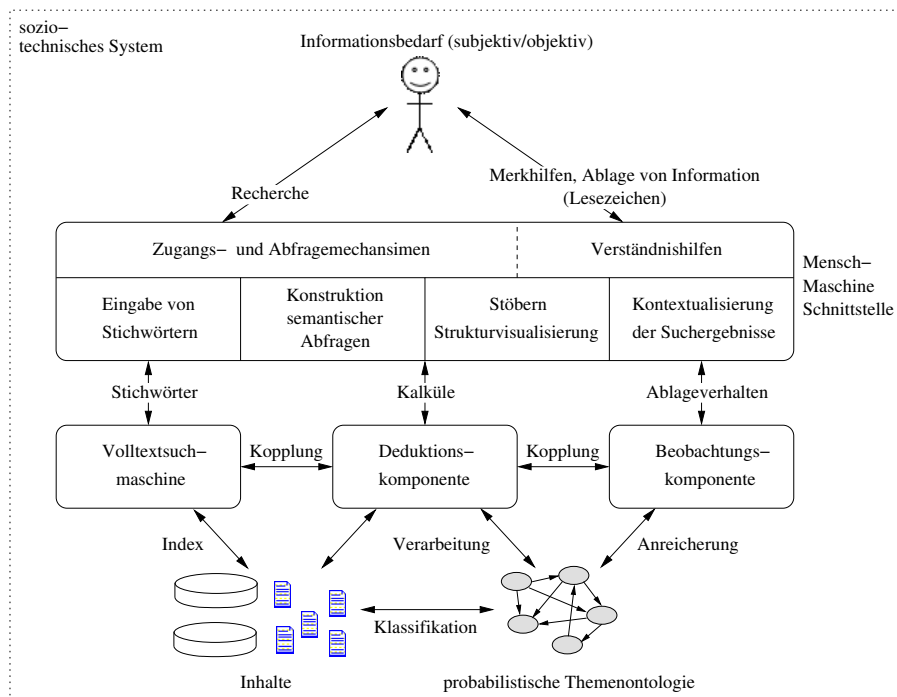


Abbildung 3.2: Integrierter Ansatz zur wissensbasierten Informationsrecherche

wählbaren Thema nach dem Prinzip der direkten Manipulation gekoppelt. Zusätzlich sind damit stichwortbasierte und semantische Abfragemöglichkeiten verbunden. Dadurch werden drei Arten von Informationszugängen geschaffen: Auf Strukturebene kann navigational durch Verbreiterung und Einengung eines Themas, sowie der Verfolgung von bedeutungsvollen Querbeziehungen gestöbert und unmittelbar die zugehörigen Inhalte in der Tabelle eingesehen und recherchiert werden. Eine Volltextsuche ermöglicht eine ungenaue Suche auf Basis von Stichwörtern. Mittels semantischer Abfragen lässt sich Information anhand von Themen und Merkmalen lokalisieren. Semantische Abfragen können visuell konstruiert werden, womit die Erlernung einer logischen Abfragesprache entfällt. Eine gleichzeitige Darstellung von Suchergebnissen und semantischer Struktur ermöglicht zudem eine Kontextualisierung der Ergebnisse in dieser Struktur. Über den Themenbezug lässt sich so interaktiv eine ungenaue Suche weiter verfeinern.

Die Grundlage der Strukturvisualisierung und der semantischen Abfragen ist die Abstraktion der suchbaren Inhalte durch eine mit Wahrscheinlichkeiten erweiterten Ontologie. Mittels Wahrscheinlichkeiten können Begriffsüberdeckungen, sowie Zugehörigkeitswerte von Inhalten zu Themen bei der Modellbildung der Ontologie und der Abfrage berücksichtigt werden. Die maschinelle Verarbeitung der ontologischen Wissensbasis erfolgt durch eine entsprechend ausgelegte Deduktionskomponente. Eine Stichwortsuchmaschine ist mit der Deduktionskomponente derart gekoppelt, dass sich zu einzelnen Suchergebnissen sämtliche semantischen Merkmale aus der Ontologie und andersherum abfragen lassen.

Innerhalb der Nutzerschnittstelle ist, wie in vielen anderen Informationssystemen, die Möglichkeit vorgesehen, Merkhilfen in Form von Lesezeichen zu erstellen. Diese Lesezeichen ermöglichen eine persönliche Klassifikation öfters benötigter, relevanter Inhalte. Zudem stellt die Erzeugung von Lesezeichen seitens der Nutzer Kontexte dar, in denen diese ihr Wissen über vorhandene Inhalte explizieren. Zur systemischen Wissensbildung existiert daher eine weitere Komponente, die das Ablageverhalten von Information der Nutzer beobachtet. Durch die Analyse dieser Lesezeichen werden neue Begriffe und Relationen der Ontologie gelernt beziehungsweise generiert.

3.2.1 Argumentation des Ansatzes

Den oben formulierten Anforderungen kann auf unterschiedlichen Arten entsprochen werden. Ein wichtiger Aspekt dabei ist die Aufgabenverteilung zwischen Mensch und Rechnersystem. Dabei gibt es zwei extreme Positionen: Die eine Position sieht den Menschen als die alleinige Instanz, die durch Verarbeitung von Information neue Erkenntnisse beziehungsweise Wissen schlussfolgert. Seine nächsten Aktionen werden darauf hin ausgerichtet. Der Rechner führt einzig und allein die eingegebenen Kommandos des Benutzers aus. Aus diesem Blickwinkel wird versucht, die Nutzungsschnittstelle so zu optimieren, dass ein schnelleres und umfassenderes Verständnis des Informationsraums und des Informationsbedarfes eintritt.

Die andere Position versucht möglichst viel Wissen über die Zusammenhänge und Strukturen des Informationsraums, sowie menschliche Suchstrategien auf das System zu übertragen. Damit wird versucht, natürliches Kommunikationsverhalten nachzuahmen. Der Rechner optimiert die menschliche Suchanfrage, zum Beispiel unter Beachtung des Arbeitskontextes und versucht, durch intelligente Suchalgorithmen bessere Ergebnisse zu erzielen. Dies kann zum Beispiel durch Techniken der Verarbeitung der natürlichen Sprache geschehen. Bei Unklarheiten werden Rückfragen gestellt und so im Dialog einigermaßen präzise Antworten geliefert.

Die beiden Positionen schließen sich aber gegenseitig nicht aus, sie sind in gewisser Hinsicht sogar orthogonal zueinander und können sich ergänzen. So ist es durchaus denkbar, dass der Rechner durch Optimierung der Suchanfrage bessere Ergebnisse findet und durch entsprechende Gestaltung der Nutzerschnittstelle möglichst viel Information sichtbar gemacht wird, die das Verständnis für Informationsraum und -bedarf erhöht. Ein alleiniges Vertrauen auf maschinelles Verständnis und "intelligente" Suchalgorithmen scheint allerdings wenig Erfolg versprechend. Durch den Kenntnismangel des objektiven Informationsbedarfs kann dieser kaum durch ein Rechnersystem aufgedeckt und ausgeglichen werden, da es auf die Eingaben des Nutzers angewiesen ist und dadurch wenig andere Information über den Informationsstand des Nutzers hat. Allenfalls die Korrektur einfacher Fehler, wie etwa Rechtschreibfehler, oder Anfrageerweiterung sind denkbar. Die Behebung einzelner, subjektbezogener Verständnisprobleme ist ein hartes Problem der Pädagogik, dessen Beschreibung mit formalen Mitteln mit den heutigen Kenntnissen fast aussichtslos erscheint.

Der hier verfolgte Ansatz geht von den Anforderungen der oben geschilderten Suchproblematik aus und legt den Fokus auf eine effizientere Nutzerschnittstelle. Im Besonderen werden effektive Zugangsmechanismen bereitgestellt, welche Suchaufgaben von der Beantwortung einer spezifischen Frage bis zu der allgemeinen Recherche eines Themengebiets unterstützen und Instrumente zur Förderung des Verständnisses von Informationsraum und des Informationsbedürfnis angeboten. Diese Instrumente werden zusammen mit den Zugangsmechanismen und der Suchergebnisdarstellung in einer einheitlichen Nutzerschnittstelle integriert, um mit möglichst wenig Kontextwechseln auszukommen und damit den Nutzer mental möglichst wenig zu belasten.

Der Forderung nach einer Verständnishilfe des zugrunde liegenden Informationsraums wird durch eine neuartige, leistungsfähige Visualisierungs- und Explorationstechnik entsprochen. Mit deren Hilfe kann der Informationsraum in einer kompakten Form dargestellt und in seiner Gesamtheit navigatorisch erkundet werden. Eine Visualisierung hat zusätzlich den Vorteil, dass sie als Bild wahrgenommen wird. Im Gegensatz zur seriellen Verarbeitung von sprachlicher und textueller Kommunikation erfolgt die menschliche Wahrnehmung von Bildern parallel und damit schneller. Durch die schnellere Wahrnehmung wird demnach die Bandbreite der Kommunikation erhöht (vgl. Abschnitt 2.6, S. 57).

Gemäß dem Referenzmodell der Informationsvisualisierung (vgl. ebenfalls Abschnitt 2.6, S. 57) erfolgt eine Datentransformation von den Rohdaten zu strukturierten Daten. Diese Datentransformation hat in der Regel auch eine Verdichtungsfunktion. Im Anwendungsfall entsprechen die Rohdaten dem Informationsraum. Als Transformation wird die semantische Abstraktion des Informationsraums zu einem hierarchischen und multifacettierten Themennetz herangezogen. Dies bedeutet die Zusammenfassung einzelner Inhalte und deren Zuordnung zu übergeordneten Themengebieten. Dieser Ansatz entspricht weitestgehend der bekannten Systematik, aus dem Bibliothekswesen bekannten beziehungsweise des elektronischen Verzeichnisses zur Strukturie-

zung einer Wissensdomäne. Solche Klassifikationen sind meist streng hierarchisch, was bedeutet, dass Inhalte jeweils nur einem einzigen Thema zugeordnet werden dürfen. Viele Dokumente behandeln allerdings mehrere Themengebiete in unterschiedlicher Ausprägung, was sowohl eine alleinige Einordnung in ein Themengebiet als auch das Auffinden erschwert. Auch können spezielle Themengebiete oft nur schwer direkt in einem einzigen, allgemeineren Thema klassifiziert werden, was zu unintuitiven Begriffshierarchien führt (vgl. Abschnitt 2.3.1, S. 31). Zudem trägt eine Hierarchie der vernetzten Natur von Themengebieten keine Rechnung. Verschieden typisierte Querbeziehungen zwischen Themen helfen aber bei dem Auffinden entsprechend verwandter Themen und von zusätzlicher Information.

Als eine Erweiterung des Themennetzes wird die Möglichkeit vorgesehen, Knoten und Kanten des Netzes mit probabilistischen Vertrauenswerten gewichten zu können. Mit Beziehungsgewichten kann der Grad, mit dem ein bestimmtes Thema durch ein jeweiliges Dokument behandelt wird oder Sub- beziehungsweise Superthema eines anderen Themas ist, näher spezifiziert werden. Auch Querbeziehungen zwischen Themen lassen sich granularer ausmodellieren. Über Knotengewichte kann die allgemeine Relevanz von Inhalten, aber auch von Themen berücksichtigt werden. Dies dient zur Rangfolgesortierung der Ergebnistabelle und erlaubt eine feinere Kontrolle der semantischen Suche.

Die Visualisierung eines solchen Themennetzes erlaubt gleich zu Beginn einer Recherchesitzung die Kommunikation mit dem System sowohl bezüglich der Struktur als auch bezüglich der thematischen Bedeutung der recherchierbaren Inhalte. Durch die verdichtete Darstellung und der Explorationsmöglichkeit auf Ebene der Themengebiete lässt sich so schnell ein Überblick gewinnen. Zusätzlich erfolgen eine Kopplung der Visualisierung mit der Ergebnispräsentation nach dem Prinzip der direkten Manipulation und eine Verbindung mit den Zugangsmechanismen. Auf diese Weise kann eine Vielzahl von Suchaufgaben bearbeitet und Einblicke in die Struktur der Ergebnismenge erreicht werden. Auf Grund der Klassifikation der Inhalte können diese themenbasiert aufgefunden werden. Zusätzlich kann auch eine thematische Kontextualisierung von Stichwortsuchergebnissen erfolgen, in dem diejenigen Themengebiete in der Visualisierung markiert werden, in denen Treffer erzielt wurden. Dies entspricht einer visuellen Auflösung von Mehrdeutigkeiten der Suchergebnisse. Mehrdeutigkeiten lassen sich aufgrund der Komplexität der Sprache bei einer Stichwortsuche nicht vermeiden (vgl. Abschnitt 2.4.6, S. 40). Über den Themenbezug werden die Inhalte gefunden, welche bestimmte Stichwörter enthalten, aber auch zu einem bestimmten Thema gehören. Dies sei an einem Beispiel verdeutlicht: Bei der Recherche in einem Intranet eines Automobilherstellers nach der Kommunikationsnorm von CAN-Bus Systemen mit dem Stichwort "CAN-Bus" werden zusätzlich zum Gesuchten Inhalte mit Systembeschreibungen und Baukomponenten gefunden. Mit dem Bezug zum Thema "Normen" findet sich das Gewünschte ohne aufwändiges Variieren der Suchterme.

Für die maschinelle Verarbeitung des beschriebenen, hierarchischen Themennetzes erfolgt dessen logische Formalisierung in einer Domänenontologie. Ontologien bieten einen maschinenverstehbaren Repräsentations- und Austauschformalismus begrifflichen Wissens, sowie die Möglichkeit zur automatischen Schlussfolgerung von impliziten Beziehungen zwischen semantischen Strukturelementen (vgl. Abschnitt 2.5.1, S. 42). Diese impliziten Beziehungen dienen als zusätzliche visuelle Navigationshilfe zur Verfolgung von Querbeziehungen. Auch ermöglicht eine Ontologie als Modellierungsformalismus die einfache Integration von nicht textuellen Inhalten wie Produkt- und Personaldaten, elektronischen Diensten und sonstigen Datenbanken. Es lassen sich ebenfalls zusätzliche Metadaten, die Inhalte auszeichnen, wie Autor, Erscheinungsjahr oder Bearbeiter einbetten. Diese werden über dieselbe Nutzerschnittstelle recherchierbar. Aufgrund der logischen Struktur der Ontologie bieten sich neben dem themen- und stichwortbasierten Zugang die präzisen Instrumente einer semantischen Suche an. Über die Verknüpfung von Themenklassen, Metadaten und bedeutungsvollen Beziehungen zwischen Themengebieten können recherchierbare Elemente gezielt aufgefunden werden. Die Ontologie dient daher zusammen mit einer entsprechenden Deduktionskomponente direkt zur Informationsrecherche und nicht, wie in herkömmlichen so genannten semantischen Suchmaschinen, zur Anfrageverfeinerung einer Stichwortsuche.

Die Verwendung eines ontologischen Themennetzes als Grundlage für die Visualisierung und Erschließung eines Informationsraums ist trotz der genannten Vorteile nicht unproblematisch. So

existieren auf der visuellen Ebene keine geeigneten Werkzeuge zur Darstellung vernetzter Begriffsstrukturen (vgl. Abschnitt 2.6.1, S. 59) im Hinblick auf unterschiedliche Nutzeraufgaben wie die Suche spezifischer Knoten und Kanten, das Entdecken von Beziehungen zwischen beliebigen Knoten oder die Erfassung aller Beziehungen eines bestimmten Knotens. Ferner finden sich auch keine zweckdienlichen visuellen Ansätze zur Konstruktion semantischer Suchanfragen. Die Eingabe einer Anfrage in einer Logiksprache ist für den durchschnittlichen Nutzer kaum zumutbar.

Auf Ebene der ontologischen Modellbildung stellt sich die grundsätzliche Frage, wie ein hierarchisches und multifacettiertes Themennetzwerk in der entsprechenden Qualität und unter Zustimmung aller Nutzer möglichst automatisch generiert werden kann (vgl. Abschnitt 2.5.1.4, S. 49). Zudem lassen sich bei den vom W3C als Standard vorgeschlagenen Spezifikationsprachen (DAML+OIL, OWL) mit ihren Sprachkonstrukten keine Beziehungs- und Knotengewichte berücksichtigen.

Diese Problematiken bedingen die Entwicklung neuer Ansätze zur Erstellung, Visualisierung und Abfrage ontologischer Begriffsnetze, sowie bei den Modellierungsprimitiven, welche eine Gewichtung von Beziehungen erlauben müssen.

Im Rahmen dieser Arbeit wird eine neue Visualisierungstechnik vorgeschlagen. Sie beruht auf der Darstellung einer Adjazenzmatrix, in der die Knoten des Netzes auf den Achsen und Relationen in den Zellen der Matrix dargestellt werden. Darauf aufbauend wird eine Möglichkeit zur visuellen Konstruktion von semantischen Suchanfragen entwickelt, mit der alle Inhalte des Informationsraums erreicht werden können. Diese Visualisierungs- und Abfragetechniken werden zusammen mit der Möglichkeit einer Stichwortsuche und der Kontextualisierung von Suchergebnissen in einer einheitlichen Nutzerschnittstelle integriert.

Zur Formalisierung der Ontologie wird die Sprache OWL Full (vgl. Abschnitt 2.5.1.3, S. 47) herangezogen, die mit Modellierungsprimitiven zur Berücksichtigung von Beziehungs- und Knotengewichten erweitert wird. Von OWL als W3C-Standard wird erwartet, eine größere Verbreitung zu finden. Zudem enthält nur OWL Full die zur Formalisierung der geforderten Ontologie benötigte formale Semantik. Mit der Erweiterung von OWL durch Beziehungsgewichte können Themenüberdeckungen und -zugehörigkeiten, sowie Relevanzen von Inhalten und Themen berücksichtigt werden.

Zur Modellbildung wird das in Abschnitt 2.5.1.4, S. 49 genannte Rahmenwerk zur Generierung von Ontologien² um die Möglichkeit erweitert, Kommunikationsprozesse zwischen Nutzer(n) und System, sowie den daraus entstandenen Artefakten analysieren und in die Ontologie einarbeiten zu können. Solche Artefakte sind explizit klassifizierte Lesezeichen und Ordnungsstrukturen (Verzeichnisse) des Arbeitsplatzes, die maschinell erfasst werden können. Diese haben den Vorteil, auch Begriffe zu enthalten, die nicht notwendigerweise in den Dokumenten auftreten und die damit von Textanalysewerkzeugen nicht gefunden werden können. Sie stellen eine persönliche Sicht des jeweiligen Nutzers mit seiner Begriffsbildung auf den Informationsraum dar³. Eine Studie von Boardman⁴ belegt, dass immerhin 50% von avancierten Nutzern ihre Arbeitsplatzressourcen wie Dateisystem, Mailprogramme und Browserlesezeichen in eine hochstrukturierte Form mit mehreren Hierarchieebenen bringen. Aufgrund dieser hierarchischen Ordnung können Ablagestrukturen zur Klassifikation verwendet werden. Durch Überlagerung über alle Nutzer hinweg lassen sich statistisch Begriffsgemeinsamkeiten und -hierarchien berechnen. Vertrauenswerte dienen als Gewicht der Beziehungen. Eine gemeinsame Begriffsbildung wird gefördert, da alle Nutzer entsprechend berücksichtigt werden.

3.3 Semantisches Modell der ontologischen Wissensbasis

Die systemische Wissensbasis auf Basis einer themenorientierten Domänenontologie ist ein grundlegender Baustein im gewählten Ansatz. Sie stellt die Abbildung des zugrunde liegenden

²MAEDCHE/STAAB: *Mining Ontologies from Text*, 2000

³NOVAK/KUNZ/WURST: *Entdeckung und Nutzbarmachung von stillem Wissen in heterogenen Expertengemeinschaften*, 2003

⁴BOARDMAN: *Multiple Hierarchies in User Workspace*, 2001

Informationsraum auf ein logisches Modell von dessen semantischer Struktur dar. Im Minimalfall wird dieser durch eine Dokumentsammlung jedwelchen Formats gebildet, kann aber aus anderen oder zusätzlichen Trägern von Inhalten bestehen. Zusätzliche Inhalte sind zum Beispiel Organigramme, Personalprofile oder Produktdaten.

Die Transformation des Informationsraums in ein logisches Modell erfolgt anhand bedeutungsvoller Merkmale der einzelnen Inhalte und deren übergelagerter Themenstruktur. Damit erfüllt die Wissensbasis zwei wesentliche Funktionalitäten: Zum einen wird sie aufgrund ihrer Verdichtungseigenschaft als Grundlage zur Visualisierung der semantischen Struktur des Informationsraums herangezogen. Zum anderen dient sie zur Beantwortung von logischen Anfragen, die eine Lokalisierung einzelner Elemente des Informationsraums anhand semantischer Merkmale ermöglichen.

Dieser Sachverhalt erfordert große Sorgfalt bei der Gestaltung eines geeigneten semantischen Modells, in dem die bedeutsamen Eigenschaften von Inhalten integrierbar sind. Dieses semantische Modell muss vor allem formal korrekt aber auch leicht auf eine Visualisierung abbildbar sein. Eine weitere Anforderung ist eine intuitive Strukturierung, da es dem Nutzer direkt präsentiert wird. Obwohl die Modellierung von Dokumenten, welche innerhalb von Themengruppen klassifiziert sind, ein gängiges Gestaltungsproblem wissensbasierter Systeme bilden, finden sich in der Literatur keine Ansätze, die sowohl dem aktuellen Problem gerecht werden als auch formal einwandfrei sind.

3.3.1 Anwendung und Anforderungen

Dokumente können nach verschiedenen Kriterien erschlossen werden. Es wird dabei zwischen formalen und inhaltlichen Merkmalen unterschieden (vgl. Abschnitt 2.3, S. 30). Des Weiteren zeichnet sich ein Dokument durch seinen Typ, seinen Art und sein Format aus. All diese Kennzeichen werden als Metadaten bezeichnet.

Die Formalerschließung oder alphabetische Katalogisierung bezeichnet die Beschreibung eines Dokuments anhand dessen formaler Gegebenheiten. Dies sind unter anderem Kennzeichen wie der Titel des Dokuments, der Name des Verfassers, die ISBN⁵-Nummer, der Verlag oder das Publikationsjahr. Die Sach- oder Inhalterschließung beschreibt ex verbis den Inhalt eines Dokuments. Diese erfolgt neben der Vergabe von Schlag- und Stichwörtern mittels einer so genannten Systematik anhand einer Klassifikation. Je nach Klassifikationssystem ergeben sich dadurch Monohierarchien oder Polyhierarchien. Sie unterscheiden sich dadurch, ob es erlaubt ist, einzelnen Klassen beziehungsweise Dokumenten nur eine oder mehrere Oberklassen zuzuordnen.

3.3.1.1 Anwendung

Eine gebräuchliche Klassifikationsmöglichkeit ist hierbei die Zuordnung von Dokumenten zu allgemeinen Themengebieten. Die einzelnen Themengebiete lassen sich wiederum in einer Themenhierarchie anordnen. Weitere Klassifikationsmöglichkeiten ergeben sich aus dem Typ, der Art und dem Format eines Dokuments. Der Dokumenttyp bezeichnet eine Kombination aus inhaltlich-strukturellen Merkmalen. So kann zum Beispiel nach Korrespondenz, Bericht oder Webseite etc. unterschieden werden. Die Dokumentart beschreibt die äußerliche Form eines Dokuments, wie zum Beispiel Text, Bild, Tabelle, während das Dokumentformat die technische Aufbereitung eines Dokumentes bezeichnet. Elektronische Dokumente können zum Beispiel als Hypertext (HTML) oder im PDF-Format vorliegen.

Während es bei der Modellierung von Dokumentkollektionen verhältnismäßig einfach ist, die formalen Kennzeichen, sowie Art, Typ und Format eines Dokuments zu erheben - meist lassen sich diese automatisch extrahieren - ist es deutlich schwieriger eine übergeordnete Themenklassifikation zu erstellen und Dokumente darin einzuordnen. Dies sei am nachfolgenden Beispiel verdeutlicht:

⁵International Standard Book Number

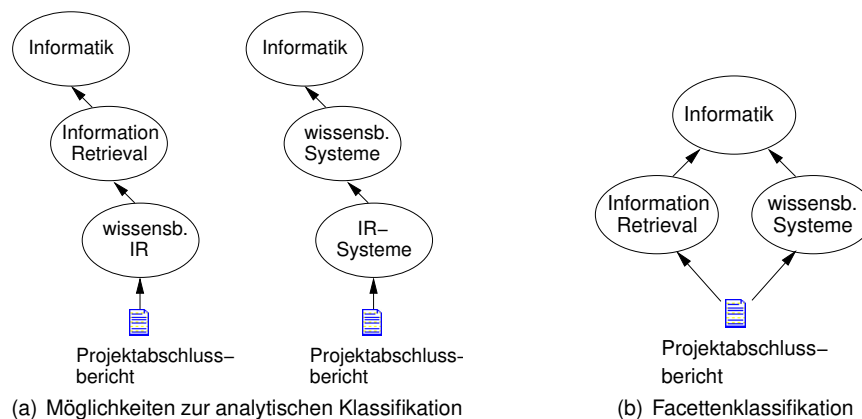


Abbildung 3.3: Analytisch Klassifikation vs. Facettenklassifikation

Beispiel Ein gängiger Dokumenttyp im Umfeld der Forschung und Entwicklung ist der des Projektberichts. Innerhalb des Forschungsprojekts "wissensbasiertes Information Retrieval" wurde ein entsprechendes System entwickelt, das wissens- und statistikbasierte Methoden des IR koppelt. Im Projektabschlussbericht werden daher unter anderem die übergeordneten Themen "Information Retrieval" zu 20% und "wissensbasierte Systeme" zu 60% behandelt. Beides sind wiederum Teilgebiete der "Informatik".

Um in der Ontologie von diesem Beispiel das Gebiet des "wissensbasierten Information Retrievals" adäquat beschreiben zu können, müssten in der üblichen analytischen Klassifikation entsprechende Unterklassen zu "Information Retrieval" und "wissensbasierte Systeme" angelegt werden (vgl. Abbildung 3.3). Der Projektabschlussbericht würde dann entweder in die Klasse "wissensbasiertes IR" oder in "IR-Systeme" der zugehörigen Oberkategorien klassifiziert.

Mit Hilfe einer polyhierarchischen Facettenklassifikation ließe sich der Projektbericht jedoch sowohl unter dem Aspekt des "Information Retrievals" als auch unter dem Blickwinkel der "wissensbasierten Systeme" einordnen. Durch Bereitstellung dieser Facetten vereinfacht sich die Systematik und wird intuitiver.

Wird eine solche Systematik um Querbeziehungen erweitert, entsteht dadurch ein multifacettiertes Begriffsnetz. Mit typisierten Querbeziehungen lässt sich auf Klassenebene der Sachverhalt modellieren, dass zum Beispiel "IR"-Techniken "wissensbasierte Systeme" "verwenden" können. Auf Instanzebene könnte dann der Projektbericht mit den tatsächlich verwendeten wissensbasierten und IR-Techniken assoziiert werden. Zusätzlich ist es wünschenswert den Grad, mit dem entsprechende Themen behandelt werden, im Modell berücksichtigen zu können.

3.3.1.2 Anforderungen

Alle oben genannten Metadaten sollten in einem entsprechenden semantischen Modell integrierbar und in seiner Spezifikationsprache ausdrucksfähig sein. Aufgrund dieses Sachverhalts und mittels allgemein gängiger ingenieurmäßiger Praktiken lassen sich die folgenden Anforderungen an ein solches semantisches Modell und seine Spezifikationsprache formulieren:

Forderung 1: Formale und eindeutige Semantik Die Hauptmotivation zur Bevorzugung von Ontologien gegenüber anderen begrifflichen Modellierungsansätzen (Themenkarten, etc) ist die Anreicherung von Daten mit maschinenverständlicher Bedeutung. Dazu müssen alle bereitgestellten Sprachkonstrukte im Hinblick auf ihre Semantik formal definiert sein. Zusätzlich zur formalen Definition darf die Semantik einzelner Konstrukte keine Mehrdeutigkeiten aufweisen damit sie nur eine einzige Interpretationsmöglichkeit zulassen. Zum Beispiel sollte die Spezifikationsprache eines semantischen Modells definieren, dass eine Klasse als eine Menge von Individuen mit gleichen Merkmalen aufzufassen ist.

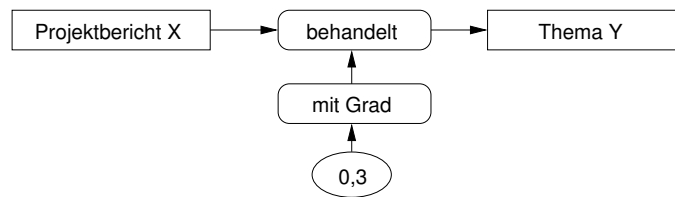


Abbildung 3.4: Beispiel einer Reifikation

Forderung 2: Formale Korrektheit Ergänzend zu der Forderung der formalen Semantik an eine Spezifikationssprache darf ein damit gebildetes semantisches Modell keine formalen Unsauberkeiten aufweisen. Zum Beispiel fordern Motik et al.⁶ von einem Modellierungsansatz die Möglichkeit einer Metamodellierung. Dies bedeutet, dass Klassen auch als Instanzen anderer Klassen, allerdings unter einer anderen Interpretation, angesehen werden können. Als Beispiel für diese Notwendigkeit führen sie die Modellierung der Beziehungen zwischen den Klassen Tierarten, Affen und den individuellen Affen in einem Bilderverwaltungssystem an: Eine natürliche Art der Modellierung wäre es zu sagen, es gäbe eine Klasse Tierarten, welche alle Tierarten, so auch die Art der Affen als Instanz beinhaltet. Jedoch kann die Art der Affen auch als Klasse aufgefasst werden, da sie alle individuellen Affen beinhaltet. Um diesen Widerspruch aufzulösen, ließe sich die Art der Affen auch als Unterklasse von Tierarten modellieren. Bei diesem Sachverhalt müsste allerdings die Klasse Tierart alle individuellen Affen als Instanz besitzen. Das ist eindeutig falsch.

Eine Metamodellierung führt mehrerer Interpretationsmöglichkeiten von Sachverhalten aus verschiedenen Blickwinkeln ein. Innerhalb eines formal korrekten semantischen Modells ist dies nicht zulässig. Falls die beschriebene Situation auftritt, sollten verschiedene semantische Ebenen eingeführt werden, welche durch das betrachtete Objekt verbunden sind. In dem einen Modell lässt sich dieses Objekt als Klasse und in dem anderen als Instanz betrachten⁷.

Forderung 3: Verwendung von Standards Standards sorgen für Interoperabilität und die Erfüllung bestimmter Minimalanforderungen. Sie sind unabhängig von dem Entwicklungsmodell, das für die Implementierung dieser Standards verwendet wurde. Gerade im Hinblick auf Wiederverwendbarkeit sollte die Spezifikationssprache eines semantischen Modells innerhalb eines Standards die Implementierung aller Anforderungen ermöglichen.

Forderung 4: Modularisierung Eine gängige Praxis im Ingenieurwesen ist die Kapselung funktionaler Bauelemente in einzelne Module, die später in verschiedenen Kontexten wiederverwendbar sind. Da Ontologien für eine Wiederverwendbarkeit konzipiert sind, sollten daher auch einzelne, semantisch zusammengehörige Teile von Ontologien gekapselt werden können.

Forderung 5: Reifikation Die Möglichkeit einer Reifikation ist ein wichtiges Werkzeug zur Ausgestaltung reichhaltiger Modelle, bei der die Forderung nach formaler Korrektheit nicht verletzt wird. Reifikation bedeutet die Vergegenständlichung von Relationen. Dadurch lassen sich Relationen attributieren. Da zwei Entitäten der Ontologie eine Aussage bilden, wenn sie mit einer Relation in Beziehung stehen, können mittels Reifikation Aussagen über Aussagen getroffen werden. Im Anwendungsfall lässt sich so eine Gewichtung von Relationen in der Spezifikationssprache einführen. Zum Beispiel kann durch Reifikation der Beziehung "behandelt" in der Aussage "Der Projektbericht X behandelt Thema Y" zu "Der Projektbericht X behandelt Thema Y" durch die zusätzliche Relation "mit Grad" zu 30% relativiert werden. Abbildung 3.4 zeigt diesen Sachverhalt in einer Graphendarstellung.

⁶MOTIK/MAEDCHE/VOLZ: *A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications*, 2002

⁷WELTY/FERRUCCI: *What's in an Instance?*, 1994

Forderung 6: Lexikalische Information Informationssysteme die auf Ontologien beruhen, müssen mit lexikalischer Information umgehen können. So sind die Entitäten einer Ontologie in einer nutzergerechten Form beziehungsweise in mehreren Sprachen auszuzeichnen. Dazu bedarf es einer konsistenten Möglichkeit, mit lexikalischer Information umgehen zu können.

3.3.2 Spezifikationsprache einer probabilistischen Ontologie

Aus den oben formulierten Anforderung lässt sich unter der Vielzahl der möglichen Spezifikationsprachen (vgl. Abschnitt 2.5.1, S. 42) eine geeignete auswählen. Auf Grund der Forderung nach der Verwendung von Standards lässt sich der Suchraum auf RDF-basierte Spezifikationsprachen (vgl. Abschnitt 2.5.1.3, S. 47) eingrenzen. Diese sind vom W3C standardisiert und setzen sich nicht nur im Forschungsumfeld als Ontologierepräsentationen durch. Da sie auf der Auszeichnungssprache XML basieren, eignen sie sich gerade für den Austausch und die Wiederverwendung. Für die Beschreibung von Klassen, Relationen und Instanzen lässt sich eine Erweiterung von RDF, dessen Schemaspezifikation RDFS heranziehen. Für eine Anwendung im IR werden jedoch die ausdrückstärkeren Derivate wie DAML+OIL und OWL benötigt. Durch eine reichhaltige Syntax erlauben diese aufwändigere Klassenkonstruktionen und -Restriktionen. Zum Beispiel muss sich der Sachverhalt ausdrücken lassen, dass ein Dokument genau einen Titel besitzt. Dies ist mit RDFS nicht möglich.

DAML+OIL wie OWL bieten die Möglichkeit der Einbindung weiterer Subontologien und sind damit modularisierbar. Ebenfalls ist der Umgang mit lexikalischer Information standardisiert. Jegliche Ontologieelemente lassen sich durch mehrsprachige Beschriftung etikettieren. Jedoch ist die Reifikation von Relationen in DAML+OIL explizit untersagt⁸. Die Berücksichtigung von Wahrscheinlichkeiten müsste wie von Nottelmann und Fuhr⁹ vorgeschlagen, durch eine Erweiterung der XML-Syntax geschehen. Da dies durch eine proprietäre Erweiterung der Syntax und Semantik die Forderung nach Standardisierung verletzt, wird die Spezifikationsprache OWL¹⁰ in vollem Sprachumfang (OWL Full) herangezogen. OWL Full erlaubt die Verwendung sämtlicher RDF-Konstrukte und damit auch die Reifikation von Relationen. Dies erlaubt die Verwendung von beliebigen Deduktionskomponenten, die diese Sprache verstehen.

3.3.2.1 OWL Full

Die Repräsentationssprache OWL stellt eine formale Metasprache dar, welche auf den Ideen der Beschreibungslogiken (vgl. Abschnitt 2.5.1.2, S. 45) beruht. Die einzelnen Modellierungselemente sind in XML/RDF Syntax, als Subjekt-Prädikat-Objekt Tripel eines semantischen Graphen ausgedrückt und mit einer präzisen Semantik hinterlegt. Auf diese Weise unterscheidet sich OWL von dem inzwischen etablierten Datenformat XML. In XML sind die einzelnen Elemente in ihrer Namensgebung frei wählbar und besitzen anwendungsabhängige Bedeutung. So ist zum Beispiel das zur Darstellung von Webseiten dienende Format HTML ebenfalls in XML ausgedrückt. Ein Vorteil von OWL gegenüber anderen Spezifikationsprachen ist die Verfügbarkeit von Werkzeugen, die, vom jeweiligen Anwendungskontext unabhängig, Schlussfolgerungen treffen können. Das Beispiel in Abbildung 3.5 zeigt die in OWL Notation beschriebenen Klassen der Menschen und der Sterblichen. Die Klasse der Menschen steht dabei in einer Subklassenrelation zu der Klasse der Sterblichen und ist mit der Beschriftung "Die Menge aller Menschen" explizit in der Sprache Deutsch (xml:lang="de") etikettiert.

Im Folgenden erfolgt eine, von der tatsächlichen XML Syntax abstrahierte, formale Definition der Struktur, Syntax und Semantik der Spezifikationsprache OWL Full. Auf dieser Grundlage wird eine probabilistische Erweiterung eingeführt.

⁸HENDLER/MCGUINNESS: *The DARPA agent markup language*, 2000

⁹NOTTELMANN/FUHR: *pDAML+OIL: A probabilistic extension to DAML+OIL based on probabilistic Datalog*, 2004

¹⁰MCGUINNESS/VAN HARMELEN: *OWL Web Ontology Language Overview*, 2003

```

<owl:Class rdf:ID="Menschen">
  <rdfs:subClassOf rdf:resource="#Sterblich" />
  <rdfs:label xml:lang="de">Die Menge aller Menschen</rdfs:label>
</owl:Class>

```

Abbildung 3.5: Klassenbeschreibung in OWL

Struktur

Eine Ontologie enthält eine begriffliche Beschreibung von Klassen, Individuen und deren Zusammenhang in einer Anwendungsdomäne. Üblicherweise wird eine Ontologie in eine *Terminologiekomponente* (TBox) und eine *Assertionalkomponente* (ABox) unterteilt (vgl. 2.5.1.2, S. 45). In der Terminologiekomponente wird das begriffliche Wissen über den betrachteten Anwendungsbereich, das Vokabular hinterlegt, wohingegen die Assertionalkomponente Aussagen über spezielle Situationen im Anwendungsbereich enthält. Ferner können bestehende Ontologien eingebunden werden. Die Struktur einer Ontologie definiert sich dadurch wie folgt:

Definition 3.1: Ontologie: Eine Ontologie ist ein 3-Tupel $\mathcal{O} = (\mathcal{T}, \mathcal{A}, \mathcal{O}_{INC})$ wobei gilt:

- \mathcal{T} ist die Terminologiekomponente
- \mathcal{A} ist die Assertionalkomponente
- \mathcal{O}_{INC} ist die endliche Menge der eingebundenen Ontologien

In der Terminologiekomponente wird das Vokabular eines Gegenstandsbereichs mit Hilfe von definierten *Begriffsbeschreibungen* hinterlegt. Neue Begriffsbeschreibungen lassen sich durch logische Verkettung bestehender Begriffsbeschreibungen und Attributen, Randbedingungen, sowie Beziehungen untereinander definieren. Die kleinste begriffliche Einheit bilden dabei atomare *Begriffe*, welche sich durch sich selbst definieren beziehungsweise nicht durch Verkettung gebildet wurden. Namentlich benannte Beziehungen zu anderen Begriffen und Attributen werden als Rollen bezeichnet. In diesem Sinne lässt sich ein solcherart definierter Begriff auch als semantische Beschreibung einer Klasse in einem Anwendungsbereich betrachten.

Definition 3.2: Terminologiekomponente: Eine Terminologiekomponente ist ein 3-Tupel $\mathcal{T} = (\mathcal{D}, \mathcal{C}, \mathcal{A}_{\mathcal{T}})$ wobei gilt:

- \mathcal{D} ist eine Menge von Datentypen
- \mathcal{C} ist eine nichtleere und endliche Menge aller Begriffsbeschreibungen
- $\mathcal{A}_{\mathcal{T}}$ ist eine endliche Menge von terminologischen Axiomen

In der Assertionalkomponente wird konkretes, situationenbezogenes Wissen modelliert. Konkrete Objekte werden benannt und als Individuen in der Assertionalkomponente eingeführt. Ihre Eigenschaften werden durch Zuordnung von Individuen zu Begriffen (Instantiierung) festgelegt und Beziehungen zwischen Objekten durch die Verknüpfung der zugehörigen Individuen durch entsprechende Rollen dargestellt (binäre Relationen).

Definition 3.3: Assertionalkomponente: Eine Assertionalkomponente ist ein 3-Tupel $\mathcal{A} = (\mathcal{I}, \mathcal{A}_{\mathcal{A}})$ wobei gilt:

- \mathcal{I} ist die endliche Menge aller Individuen (beziehungsweise die Instanzen von \mathcal{C} der Terminologiekomponente \mathcal{T})
- $\mathcal{A}_{\mathcal{A}}$ ist die endliche Menge von Assertionen bezüglich der Menge aller Individuen \mathcal{I}

Syntax

Die Basiselemente der Spezifikationssprache OWL sind Datentypen \mathbf{D} und Begriffsbeschreibungen \mathbf{C} (einstellige Prädikate), Rollen \mathbf{R} (zweistellige Prädikate), sowie terminologische Axiome \mathbf{A}_T und Assertionen \mathbf{A}_I . Begriffsbeschreibungen \mathbf{C} können atomar (\mathbf{C}_A) oder komposit (\mathbf{C}_K) sein. Dabei werden komposite Begriffsbeschreibungen aus existierenden durch weitere Begriffsbeschreibungen definiert. Atomare Begriffe besitzen weder Attribute noch Beziehungen zu anderen Begriffsklassen. Zum Beispiel stellen die Begriffsbeschreibungen "Person", "Mann", "Frau" atomare Begriffe dar. Jede dieser Begriffsbeschreibungen stellt eine Menge von Individuen \mathbf{I} dar. Alle Rollen \mathbf{R} sind geordnete Paare von Individuen. Dabei wird zwischen abstrakten ($\mathbf{R}_A \in \mathbf{R}$) und funktionalen Rollen ($\mathbf{R}_F \in \mathbf{R}$) unterschieden. Abstrakte Rollen stellen Beziehungen zwischen Begriffsbeschreibungen dar, während funktionale Rollen Beziehungen zwischen Begriffsklassen und Attributen beziehungsweise Datentypen repräsentieren.

Definition 3.4: Begriffsbeschreibungen: Sind \mathbf{C}_A , \mathbf{R}_A , \mathbf{R}_F , \mathbf{I} die nichtleeren, endlichen und disjunkten Mengen von atomaren Begriffen, abstrakten und funktionalen Rollen, sowie Individuen, so definieren sich Begriffsbeschreibungen rekursiv und durch Induktion wie folgt:

- jeder atomare Begriff $\mathbf{C}_A \in \mathbf{C}$ ist eine Begriffsbeschreibung \mathbf{C}
- sind $i_1, i_2 \dots i_d \in \mathbf{I}$ Individuen aus \mathbf{I} , so ist $\{i_1, i_2 \dots i_d\}$ eine Begriffsbeschreibung
- sind $\mathbf{C} \in \mathbf{C}$ und $\mathbf{D} \in \mathbf{C}$ Begriffe aus \mathbf{C} , so sind ebenfalls

$\mathbf{C} \sqcap \mathbf{D}$	(Konjunktion)
$\mathbf{C} \sqcup \mathbf{D}$	(Disjunktion)
$\neg \mathbf{C}$	(Negation)

 Begriffsbeschreibungen.
- ist $\mathbf{C} \in \mathbf{C}$ ein Begriff aus \mathbf{C} , $\mathbf{R} \in \mathbf{R}_A$ eine abstrakte Rolle aus \mathbf{R}_A und d eine nicht negative ganz Zahl, so sind auch die folgenden Ausdrücke Begriffsbeschreibungen:

$\exists \mathbf{R}.\mathbf{C}$	(Existenzrestriktion)
$\forall \mathbf{R}.\mathbf{C}$	(Werterestriktion)
$\exists_{\geq n} \mathbf{R}$	(Maximum-Kardinalitätsrestriktion)
$\exists_{\leq n} \mathbf{R}$	(Minimum-Kardinalitätsrestriktion)
$\exists_{\geq n} \mathbf{R}.\mathbf{C}$	(Qualifizierende Maximum-Kardinalitätsrestriktion)
$\exists_{\leq n} \mathbf{R}.\mathbf{C}$	(Qualifizierende Minimum-Kardinalitätsrestriktion)

 Exakt qualifizierende Kardinalitätsrestriktionen $\exists_{\geq n} \mathbf{R} \sqcap \exists_{\leq n} \mathbf{R}$ (bzw. $\exists_{\geq n} \mathbf{R}.\mathbf{C} \sqcap \exists_{\leq n} \mathbf{R}.\mathbf{C}$) werden mit $\exists_{=n} \mathbf{R}$ (bzw. $\exists_{=n} \mathbf{R}.\mathbf{C}$) abgekürzt.
- ist $\mathbf{C} \in \mathbf{C}$ ein Begriff aus \mathbf{C} , $\mathbf{T} \in \mathbf{R}_F$ eine funktionale Rolle aus \mathbf{R}_F , n eine nicht negative ganz Zahl und $\mathbf{d} \in \mathbf{D}$ ein konkreter Datentyp aus \mathbf{D} , so sind auch

$\exists \mathbf{T}.\mathbf{d}$	(Datentyp Existenzrestriktion)
$\forall \mathbf{T}.\mathbf{d}$	(Datentyp Werterestriktion)
$\exists_{\geq n} \mathbf{T}.\mathbf{C}$	(Datentyp qualifizierende Maximum-Kardinalitätsrestriktion)
$\exists_{\leq n} \mathbf{T}.\mathbf{C}$	(Datentyp qualifizierende Minimum-Kardinalitätsrestriktion)

 Begriffe. Exakt qualifizierende Datentyp Kardinalitätsrestriktionen $\exists_{\geq n} \mathbf{T}.\mathbf{C} \sqcap \exists_{\leq n} \mathbf{T}.\mathbf{C}$ werden mit $\exists_{=n} \mathbf{T}.\mathbf{C}$ abgekürzt.
- ferner wird mit \top der universelle Begriff $\mathbf{C} \sqcup \neg \mathbf{C}$ und mit
- \perp der leere Begriff $\mathbf{C} \sqcap \neg \mathbf{C}$ abgekürzt

Terminologische Axiome sind Teil der Terminologiekomponente und drücken Annahmen bezüglich Begriffen und Rollen aus. So lassen sich Subsumptionsbeziehungen von Begriffen und Rollen modellieren. Ferner ermöglichen terminologische Axiome die Definition von inversen, symmetrischen und transitiven Beziehungstypen.

Definition 3.5: terminologische Axiome \mathbf{A}_T Sind $\mathbf{C} \in \mathbf{C}$ und $\mathbf{D} \in \mathbf{C}$ Begriffe aus \mathbf{C} , $\mathbf{R}, \mathbf{S} \in \mathbf{R}_A$ oder $\mathbf{R}, \mathbf{S} \in \mathbf{R}_F$ Rollen aus \mathbf{R} so nennen sich die Ausdrücke

- $\mathbf{C} \sqsubseteq \mathbf{D}$ Begriffsinklusion: \mathbf{C} ist damit ein spezifischerer Begriff als \mathbf{D} .

- $R \sqsubseteq S$ Rolleninklusion: R ist damit eine spezifischere Rolle als S .
- $R^{-1} \doteq S$ Inversionsaxiom: R ist damit die inverse Rolle zu S .
- $R \circ S$ Kompositionsaxiom: R und S werden zu einer Rollenkette (Pfad) zusammengefügt.
- R^+ Transitivitätsaxiom, wobei $\forall R \in \mathbf{R}_A$: R^+ ist damit eine der transitive Abschluss der abstrakten Rolle R .
- R^S Symmetrieaxiom: R ist damit eine symmetrische Rolle.
- $\exists R \sqsubseteq C$ Wertebereich von R .
- $\top \sqsubseteq \forall R.C$ Definitionsbereich von R .

Die Äquivalenz von zwei Begriffen C, D kann durch die beiden Inklusionsaxiome $C \sqsubseteq D$ und $D \sqsubseteq C$ ausgedrückt oder durch $C \doteq D$ abgekürzt werden. Analog dazu lässt sich eine Rollenäquivalenz von R und S durch die beiden Inklusionsaxiome $R \sqsubseteq S$ und $S \sqsubseteq R$ ausdrücken beziehungsweise durch $R \doteq S$ abkürzen.

Assertionen drücken Sachverhalte bezüglich Individuen und deren Beziehungen hinsichtlich der Terminologiekomponente \mathcal{T} aus. So lässt sich eine Menge von Individuen \mathbf{I} durch Begriffe aus \mathbf{C} beschreiben und abstrakte Beziehungen, welche Individuen untereinander eingehen, definieren. Ferner lassen sich durch funktionale Rollen Individuen attributieren.

Definition 3.6: Assertionen \mathbf{A}_A Ist $C \in \mathbf{C}$ ein Begriff aus \mathbf{C} , $R \in \mathbf{R}_A$, $T \in \mathbf{R}_F$ Rollen aus \mathbf{R} , $d \in \mathbf{D}$ ein Datentyp aus \mathbf{D} und $i_1, i_2 \in \mathbf{I}$ ein Individuum aus \mathbf{I} , so nennt sich der Ausdruck

- $i_1 : C$ Begriffsassertion: Eine Begriffsassertion beschreibt ein Individuum i_1 durch den Begriff C (i_1 ist Instanz der mit C beschriebenen Klasse).
- $i_1 R i_2$ Rollenassertion, wobei $R \in \mathbf{R}_A$: Eine Rollenassertion setzt zwei Individuen i_1, i_2 durch die abstrakte Rolle R miteinander in Beziehung.
- $i_1 T d_1$ Attributsassertion, wobei $R \in \mathbf{R}_F$: Eine Attributsassertion attributiert eine Individuum i_1 , durch die funktionale Rolle R mit dem Datentyp d .

An dieser Stelle sei angemerkt, dass die Repräsentationssprache OWL ausdrucksstark genug ist, um obige Assertionen direkt in der Terminologiekomponente auszudrücken. Eine Begriffsassertion lässt sich auch durch eine Begriffsinklusion in der Form $\{i\} \sqsubseteq C$ ausdrücken und ist äquivalent zu $i_1 : C$. Ebenfalls lässt sich eine Rollenassertion oder Attributsassertion als $\exists R\{i_1\}.\{i_2\}$ beziehungsweise $\exists R\{i_1\}.\{d_1\}$ schreiben.

Semantik

Auf der semantischen Ebene werden Begriffe als Teilmengen eines Gegenstandsbereichs, Datentypen als Teilmengen eines Wertebereichs und Rollen als binäre Relationen sowohl über diesen Werte- als auch über den Gegenstandsbereich interpretiert. So genannte *Interpretationen* ordnen jedem atomaren Begriff eine Teilmenge des Gegenstandsbereichs und jedem Individuum ein Objekt des Gegenstandsbereichs zu. Ferner wird unter jeder abstrakten Rolle eine binäre Relation über den Gegenstandsbereich und unter jeder funktionalen Rolle eine binäre Relation zwischen dem Gegenstands- und dem Wertebereich verstanden. Die genaue Bedeutung eines Begriffs ergibt sich damit aus der folgenden mengentheoretischen Definition der Semantik.

Definition 3.7: Interpretation Das Paar $\mathcal{I} = (\Delta^{\mathcal{I}}, I)$ ist unter Berücksichtigung eines nichtleeren Gegenstandsbereichs $\Delta^{\mathcal{I}}$ und eines Wertebereichs $\text{dom}(\mathbf{D})$ von den Datentypen \mathbf{D} eine Interpretation einer Menge von atomaren Begriffen \mathbf{C}_A und einer Menge von Rollen \mathbf{R} . $\Delta^{\mathcal{I}}$ wird die Grundmenge von \mathcal{I} genannt und enthält alle Objekte, die in der Interpretation gegeben sind. I ist eine Funktion, die jeden atomaren Begriff C_A auf eine Teilmenge $C_A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ aus $\Delta^{\mathcal{I}}$, jedes Individuum $\{i_1\} : i_1 \in \mathbf{I}$ auf ein einziges Objekt $\{i_1^{\mathcal{I}}\} \subseteq \Delta^{\mathcal{I}}$ aus $\Delta^{\mathcal{I}}$, jede abstrakte Rolle R_A auf eine binäre Relation $R_A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ und jede funktionale Rolle R_F auf eine binäre Relation $R_F^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \text{dom}(\mathbf{D})$ abbildet. Die induktive Erweiterung von \mathcal{I} auf komposite Begriffe ist durch Tabelle 3.1 gegeben.

Beschreibung	Syntax	Semantik
universelle Begriffsklasse	\top	$\Delta^{\mathcal{I}}$
leere Begriffsklasse	\perp	\emptyset
Konjunktion	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
Disjunktion	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$
Negation	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
Existenzrestriktion	$\exists R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y : (x, y) \in R_A^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$
Werterestriktion	$\forall R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y : (x, y) \in R_A^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}$
Maximum-Kardinalitätsrestriktion	$\exists_{\geq n} R$	$\{x \in \Delta^{\mathcal{I}} \mid \ \{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R_A^{\mathcal{I}}\}\ \geq n\}$
Minimum-Kardinalitätsrestriktion	$\exists_{\leq n} R$	$\{x \in \Delta^{\mathcal{I}} \mid \ \{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R_A^{\mathcal{I}}\}\ \leq n\}$
Qualifizierende Maximum-Kardinalitätsrestriktion	$\exists_{\geq n} R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \ \{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R_A^{\mathcal{I}}\}\ \geq n \wedge y \in C^{\mathcal{I}}\}$
Qualifizierende Minimum-Kardinalitätsrestriktion	$\exists_{\leq n} R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \ \{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R_A^{\mathcal{I}}\}\ \leq n \wedge y \in C^{\mathcal{I}}\}$
Datentyp Existenzrestriktion	$\exists T.d$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y : (x, y) \in R_F^{\mathcal{I}} \wedge y \in \text{dom}(\mathbf{D})\}$
Datentyp Werterestriktion	$\forall T.d$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y : (x, y) \in R_F^{\mathcal{I}} \rightarrow y \in \text{dom}(\mathbf{D})\}$
Datentyp Qualifizierende Maximum-Kardinalitätsrestriktion	$\exists_{\geq n} T.d$	$\{x \in \Delta^{\mathcal{I}} \mid \ \{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R_F^{\mathcal{I}}\}\ \leq n \wedge y \in \text{dom}(\mathbf{D})\}$
Datentyp Qualifizierende Minimum-Kardinalitätsrestriktion	$\exists_{\leq n} T.d$	$\{x \in \Delta^{\mathcal{I}} \mid \ \{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R_F^{\mathcal{I}}\}\ \leq n \wedge y \in \text{dom}(\mathbf{D})\}$

Tabelle 3.1: Syntax und Semantik von Begriffsbeschreibungen.

Beschreibung	Erfüllbarkeitsbedingung	
Begriffsinklusion	$\mathcal{I} \models C \sqsubseteq D$	nur wenn $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
Rolleninklusion	$\mathcal{I} \models R \sqsubseteq S$	nur wenn $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
Inversionsaxiom	$\mathcal{I} \models R^{-1} \doteq S$	nur wenn $x, y \in \Delta^{\mathcal{I}} \mid (x, y) \in R^{\mathcal{I}} \wedge (y, x) \in S^{\mathcal{I}}$
Kompositionsaxiom	$\mathcal{I} \models R \circ S$	nur wenn $x, y, z \in \Delta^{\mathcal{I}} \mid \{(x, z) \mid \exists y : (x, y) \in R^{\mathcal{I}} \wedge (y, z) \in S^{\mathcal{I}}\}$
Transitivitätsaxiom	$\mathcal{I} \models R^+$	nur wenn $x, y, z \in \Delta^{\mathcal{I}} \mid \{(x, z) \mid \exists y : (x, y) \in R^{\mathcal{I}} \wedge (y, z) \in R^{\mathcal{I}}\}$
Symmetrieaxiom	$\mathcal{I} \models (R^S)$	nur wenn $x, y \in \Delta^{\mathcal{I}} \mid (x, y) \wedge (y, x) \in R^{\mathcal{I}}$

Tabelle 3.2: Erfüllbarkeitsbedingungen von Interpretationen

Eine Interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, I)$ erfüllt eine Terminologiekomponente \mathcal{T} , wenn sie alle korrespondierenden terminologischen Axiome $\mathbf{A}_{\mathcal{T}}$ aus \mathcal{T} erfüllt ($\mathcal{I} \models \mathbf{A}_{\mathcal{T}}$). In diesem Fall wird eine Interpretation \mathcal{I} auch ein Modell von \mathcal{T} genannt. Es gilt $\mathcal{I} \models \mathcal{T}$ genau dann, wenn $\mathcal{I} \models \mathbf{A}_{\mathcal{T}}$ für alle $\mathbf{A}_{\mathcal{T}} \in \mathcal{T}$ erfüllt ist. Jeder Begriff $C \in \mathcal{C}$ ist erfüllbar, wenn die zugehörige Terminologiekomponente \mathcal{T} ein Modell \mathcal{I} hat, in dem C mindestens ein Individuum $i_1 \in I$ aus I erzeugt wird, so dass $C^{\mathcal{I}} \neq \emptyset$ gilt. Anschaulich muss mit $C^{\mathcal{I}}$ mindestens ein reales Objekt aus dem Gegenstandsbereich $\Delta^{\mathcal{I}}$ beschrieben werden.

Die Semantik einer Terminologiekomponente ist durch die Menge ihrer Modelle gegeben. Eine Terminologiekomponente \mathcal{T} ist genau dann erfüllbar, wenn ein Modell von \mathcal{T} existiert. Ein terminologisches Axiom $\mathbf{A}_{\mathcal{T}}$ ist eine logische Konsequenz von \mathcal{T} , wenn jedes Modell von \mathcal{T} auch ein Modell von $\mathbf{A}_{\mathcal{T}}$ ist ($\mathcal{T} \models \mathbf{A}_{\mathcal{T}}$). Aus Tabelle 3.2 kann die Erfüllbarkeitsbedingung für ein bestimmtes terminologisches Axiom $\mathbf{A}_{\mathcal{T}}$ entnommen werden.

Es sei an dieser Stelle angemerkt, dass die Semantik von Begriffsbeschreibungen auch durch deren Übersetzung in Formeln der Prädikatenlogik erster Stufe festgelegt werden kann. Auf diese alternative Definition der Semantik soll hier aber nicht näher eingegangen werden, da sie der mengentheoretischen Definition gleichbedeutend ist.

Randbedingungen für die Einbindung von Subontologien

Innerhalb der Spezifikationsprache OWL kann eine modellierte Ontologie weitere Ontologien \mathcal{O}_{INC} einbinden. Dieser Mechanismus dient zur Modularisierung von bestimmten Teilen einer Ontologie, die in anderen Kontexten eingesetzt werden können. Dadurch lassen sich schon vorhandene Ontologien durch die Einbettung wieder verwenden. An die Modularisierung werden jedoch einige Bedingungen gestellt, da sowohl Begriffsbeschreibungen als auch deren Individuen bei der Kapselung betrachtet werden müssen¹¹. Zum Beispiel hat der Begriff `Kontinent` exakt sieben Individuen. Wird eine Ontologie der Kontinente in eine andere eingebunden, reicht es nicht aus, nur den Begriff `Kontinent` einzubinden. Auch die jeweiligen tatsächlichen Kontinente, wie Europa, müssen mit betrachtet werden, um Information über die Kontinente zu erhalten.

Folgende Randbedingungen sind an die Modularisierung von Ontologien zu stellen:

Definition 3.8: Randbedingungen_einer_Modularisierung: Schließt eine Ontologie \mathcal{O} eine weitere Ontologie \mathcal{O}_{INC} ein, so muss gelten

- $\mathbf{D}_{INC} \subseteq \mathbf{D}$: Alle eingebundenen Datentypen \mathbf{D}_{INC} sind Teilmenge aller verfügbaren Datentypen \mathbf{D} .
- $\mathbf{C}_{INC} \subseteq \mathbf{C}$: Alle eingebundenen Begriffsbeschreibungen \mathbf{C}_{INC} sind Teilmenge aller verfügbaren Datentypen \mathbf{C} .
- $\mathbf{A}_{TINC} \subseteq \mathbf{A}_{\mathcal{T}}$: Alle eingebundenen terminologischen Axiome \mathbf{A}_{TINC} sind Teilmenge aller verfügbaren terminologischen Axiome $\mathbf{A}_{\mathcal{T}}$.

¹¹McGUINNESS/VAN HARMELLEN: *OWL Web Ontology Language Overview*, 2003

- $\mathbf{A}_{A_{INC}} \subseteq \mathbf{A}_A$: Alle eingebundenen Assertionen $\mathbf{A}_{A_{INC}}$ sind Teilmenge aller verfügbaren Assertionen \mathbf{A}_A .

Mit diesen Randbedingungen wird die strukturelle Einheit einer eingebundenen Ontologie \mathcal{O}_{INC} sichergestellt. Die einschließende Ontologie \mathcal{O} ist jedoch nur in Verbindung mit \mathcal{O}_{INC} als Einheit zu betrachten. So kann zum Beispiel die Relation $(i_1, i_{2_{INC}}) : R_{INC}$ zwischen zwei Individuen i_1 und $i_{2_{INC}}$ nur Teil der einschließenden Ontologie \mathcal{O} sein.

Inferenzdienste der Deduktionskomponente

Auf Basis der Repräsentation von begrifflichem Wissen in der Spezifikationssprache OWL ist die Aufgabe einer Deduktionskomponente die Ableitung von Schlussfolgerungen auf Grundlage dieses Wissen. Hierzu werden Konsistenz-, Subsumptions- und Instanztest als Inferenzdienste zur Verfügung gestellt. Der Konsistenztest entscheidet, ob eine Terminologiekomponente oder ein Begriff daraus erfüllbar ist und nicht im Widerspruch zu anderen steht. Bei dem Subsumptionstest wird für zwei gegebene Begriffe bestimmt, welches der Allgemeinere ist. Schließlich entscheidet der Instanztest, ob ein gegebenes Individuum oder eine gegebene Relation Mitglied einer Menge ist, die durch eine bestimmte Begriffsbeschreibung definiert ist.

Inferenzdienste, die entscheiden, ob eine Ontologie erfüllbar und damit konsistent ist, unterstützen den Anwender bei der Modellierung, da sie Widersprüche in Begriffen und Assertionen aufdecken. Liegt dann eine konsistente Ontologie vor, so erlauben es die Dienste der Subsumption und Instanz, implizites Wissen über die definierten Begriffe und Individuen in Form von Subsumptionsbeziehungen zwischen Begriffen, sowie Instanzbeziehungen zwischen Individuen und Begriffen zu folgern.

Formal sind diese Tests:

Definition 3.9: Konsistenz Eine Terminologiekomponente \mathcal{T} ist genau dann erfüllt, wenn jede Begriffsbeschreibung $C \in \mathcal{T}$ erfüllt ist. Eine Begriffsbeschreibung C ist genau dann erfüllt, wenn aus den Erfüllbarkeitsbedingungen aus Tabelle 3.2 hervorgeht, dass $\mathcal{T} \not\models C \sqsubseteq \perp$. Eine Assertionalkomponente \mathcal{A} ist genau dann von \mathcal{T} erfüllt, wenn eine Interpretation \mathcal{I} existiert, die Modell von \mathcal{T} ist.

Definition 3.10: Subsumption Eine Begriffsbeschreibung D subsumiert die Begriffsbeschreibung C in einer Terminologiekomponente \mathcal{T} , kurz $C \sqsubseteq_{\mathcal{T}} D$ genau dann, wenn $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ für alle Interpretationen \mathcal{I} (für alle Modelle \mathcal{I} von \mathcal{T}) erfüllt ist. Es muss gelten $\mathcal{T} \models C \sqsubseteq D$.

Definition 3.11: Instanz Ein Individuum $i_1 \in \mathbf{I}$ ist bezüglich \mathcal{A} und \mathcal{T} Instanz der Begriffsbeschreibung C , kurz $i_1 \in_{\mathcal{A}, \mathcal{T}} C$, genau dann, wenn $i_1^{\mathcal{I}} \in C^{\mathcal{I}}$ für alle Modelle \mathcal{I} von \mathcal{A} und \mathcal{T} gilt. Es muss gelten $\mathcal{T} \models i_1 \in C$. Eine Relation $(i_1, i_2) : R$ mit $i_1, i_2 \in \mathbf{I}$ ist genau dann Instanz der Rolle R , wenn $(i_1^{\mathcal{I}}, i_2^{\mathcal{I}}) \in R^{\mathcal{I}}$ für alle Modelle \mathcal{I} von \mathcal{A} und \mathcal{T} ist. Es muss gelten $\mathcal{T} \models (i_1, i_2) \in R$.

Die betrachtete Spezifikationssprache erlaubt die Konjunktion und Negation von Begriffsbeschreibungen. Mittels den Abkürzungen $\{i\} \sqsubseteq C$, $\exists R\{i_1\}.\{i_2\}$ und $\exists R\{i_1\}.\{d_1\}$ lassen sich die obigen Inferenzprobleme auf den Subsumptionstest reduzieren, da die Instanztests Spezialfälle der Subsumption sind. Ebenfalls ist der Konsistenztest eine Verallgemeinerung der Subsumption und kann auf diese zurückgeführt werden. Die Konsistenz einer Terminologiekomponente ist auch erfüllt, wenn $\mathcal{T} \models \perp \sqsubseteq \top$ gilt.

3.3.2.2 pOWL - eine probabilistische Erweiterung von OWL

Die Beschreibungslogiksprache OWL erlaubt die Spezifikation von Ontologien, die auf sicheren Fakten beruhen. Eine Anfrage an eine entsprechende Deduktionskomponente liefert daher auch nur exakte Treffer zurück. Zum Beispiel wird für eine Begriffsbeschreibung die Menge aller Individuen zurückgeliefert, für die diese Begriffsbeschreibung gilt. Wünschenswert für eine

IR-Anwendung wäre zusätzlich der Umgang mit unsicherem Wissen. So sollte für eine Begriffsbeschreibung zwar die Menge aller ihrer Individuen zurückgeliefert werden, zusätzlich jedoch Information, mit welchem Vertrauensgrad diese Begriffsbeschreibung für das jeweilige Individuum gilt. Ebenfalls sollte deduziert werden können, dass eine Relation zwischen zwei Individuen zwar existiert, jedoch nur mit einer gewissen Wahrscheinlichkeit auch gültig ist. Auf diese Weise lassen sich Inhalte zu einem bestimmten Themengebiet nach Relevanz sortieren beziehungsweise zu einem bestimmten Inhalt bis zu einem gewissen Grad ähnliche Inhalte auffinden.

Aus dieser Motivation heraus werden im Folgenden *probabilistische Randbedingungen*^{12,13} für terminologische Axiome und Assertionen über Begriffe und Individuen in OWL eingeführt. Die damit erweiterten *probabilistischen Begriffsbeschreibungen* stellen Wahrscheinlichkeitsverteilungen über die Menge ihrer Individuen dar. Aufgrund der probabilistischen Randbedingungen können exakte Ontologien zu probabilistischen Ontologien erweitert werden.

Struktur

Probabilistische Ontologien bestehen aus exakten Ontologien, wobei deren exakten Terminologiekomponenten um probabilistische Randbedingungen erweitert sind. Eine probabilistische Terminologiekomponente definiert sich wie folgt:

Definition 3.12: probabilistische Terminologiekomponente: Eine probabilistische Terminologiekomponente ist ein Paar $\mathcal{T}_P = (\mathcal{T}, \mathcal{D})$ wobei gilt:

- \mathcal{T} ist die exakte Terminologiekomponente.
- \mathcal{D} ist die Menge von probabilistischen Randbedingungen

Analog zu Abschnitt 3.3.2.1 ist eine probabilistische Ontologie das 3-Tupel $\mathcal{O}_P = (\mathcal{T}_P, \mathcal{A}, \mathcal{O}_{P_{INC}})$, bestehend aus einer probabilistischen Terminologiekomponente, einer Assertionalkomponente und weiteren eingebundenen probabilistischen Ontologien.

Syntax

Eine probabilistische Randbedingung hat prinzipiell die Form $(D|C)[u, o]$. C und D sind dabei Begriffsbeschreibungen aus \mathbf{C} , während die reellen Zahlen u, o im Intervall $u, o \in [0, 1]$ liegen. Die beiden reellen Zahlen u und o bestimmen das Intervall $[u, o]$, in dem die bedingte Wahrscheinlichkeit $P(D|C) = \frac{P(C \cap D)}{P(C)}$ für das Auftreten von D liegen muss, wenn C bereits eingetreten ist. Es lassen sich damit die folgenden Sachverhalte beschreiben:

Definition 3.13: probabilistische Randbedingungen: Sind $C \in \mathbf{C}$ und $D \in \mathbf{C}$ Begriffe aus \mathbf{C} , $R \in \mathbf{R}_A$ Rollen aus \mathbf{R} und $i_1, i_2 \in \mathbf{I}$, so bedeutet

- $(D|C)[u, o]$, dass ein beliebiges Individuum $i : C$ von C mit der Wahrscheinlichkeit im Intervall $[u, o]$ auch ein Individuum von D ist. D subsumiert damit C mit der Wahrscheinlichkeit im Intervall $[u, o]$.
- $(C|\{i_1\})[u, o]$, dass ein bestimmtes Individuum $i_1 : C$ von dem Begriff C mit der Wahrscheinlichkeit im Intervall $[u, o]$ beschrieben wird.
- $(\exists R.\{i_1\}|C)[u, o]$, dass ein beliebiges Individuum $i : C$ von C durch die bestimmte abstrakte Rolle $R \in \mathbf{R}_A$ und dem speziellen Individuum i_1 mit der Wahrscheinlichkeit im Intervall $[u, o]$ in Beziehung steht.
- $(\exists R.\{i_1\}|\{i_2\})[u, o]$, dass ein bestimmtes Individuum i_1 mit dem bestimmten Individuum i_2 durch die spezielle abstrakte Rolle $R \in \mathbf{R}_A$ mit der Wahrscheinlichkeit im Intervall $[u, o]$ in Beziehung steht.

¹²GIUGNO/LUKASIEWICZ: *P-SHOQ(D): A Probabilistic Extension of SHOQ(D)*, 2002.

¹³LUKASIEWICZ: *Probabilistic logic programming with conditional constraints*, 2001

Semantik

Die Semantik einer probabilistischen Ontologie ergibt sich durch Generalisierung der bereits beschriebenen Semantik der exakten Ontologien. Die exakten Interpretationen werden hierfür mit einer Wahrscheinlichkeitsverteilung über den Gegenstandsbereich erweitert. Daraus ergibt sich die Erfüllbarkeit von terminologischen Axiomen und probabilistischen Randbedingungen in einer solchen Interpretation.

Definition 3.14: probabilistische Interpretation Eine probabilistische Interpretation ist ein Paar $P_I = (\mathcal{I}, \mu)$, das unter Berücksichtigung eines Wertebereichs $\text{dom}(\mathbf{D})$ der Datentypen \mathbf{D} aus einer exakten Interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, I)$ und einer Funktion μ auf $\Delta^{\mathcal{I}}$ besteht. Die Funktion μ ist hierbei eine Abbildung $\mu : \Delta^{\mathcal{I}} \rightarrow [0, 1]$ und gibt die Wahrscheinlichkeit eines Individuums $i^I \in \Delta^{\mathcal{I}}$ so an, dass die Summe aller $\mu(i_i^I)$ eins ergibt. Es gilt $\sum_i \mu(i_i^I) = 1$.

Die Wahrscheinlichkeit $P_I(C)$ einer Begriffsbeschreibung $C \in \mathcal{C}$ unter der probabilistischen Interpretation $\mathcal{I}_P = (\mathcal{I}, \mu)$ ergibt sich damit als die Summe der Wahrscheinlichkeiten aller ihrer Individuen $i \in C$, so dass $P(C) = \sum_c \mu(i_c^I)$ mit $i_c^I \in C^I$. Für zwei Begriffsbeschreibungen C und D mit $P(C) \geq 0$ wird die bedingte Wahrscheinlichkeit $\frac{P(C \cap D)}{P(C)}$ mit $P(D|C)$ abgekürzt.

Die probabilistische Interpretation $\mathcal{P}_{\mathcal{I}} \models (C|D)[u, o]$ erfüllt eine probabilistische Randbedingung $(C|D)[u, o]$ oder nennt sich Modell derselben genau dann, wenn $P_I(D|C) \in [u, o]$ gilt. Ebenfalls erfüllt $\mathcal{P}_{\mathcal{I}} \models \mathbf{A}_{\mathbf{A}}$ ein terminologisches Axiom $\mathbf{A}_{\mathbf{A}}$ genau dann, wenn die zugehörige Interpretation $\mathcal{I} \models \mathbf{A}_{\mathbf{A}}$ das terminologische Axiom $\mathbf{A}_{\mathbf{A}}$ erfüllt.

Inferenzdienste der probabilistischen Deduktionskomponente

Neben den Konsistenz-, Subsumptions- und Instanztests einer exakten Deduktionskomponente ist die Hauptaufgabe einer probabilistischen Deduktionskomponente die Berechnung der Intervalle $[u, o]$. Ebenfalls muss der Überdeckungsgrad zwischen Begriffsbeschreibungen bestimmt werden. Auf Grundlage der probabilistischen Semantik erweitern sich die Inferenzdienste zu den folgenden Tests:

Definition 3.15: probabilistische Konsistenz Eine probabilistische Terminologiekomponente \mathcal{T}_P ist genau dann erfüllt, wenn jede Begriffsbeschreibung $C \in \mathcal{T}_P$ erfüllt ist. Eine Begriffsbeschreibung C ist genau dann erfüllt, wenn aus \mathcal{T}_P hervorgeht, dass $\mathcal{T}_P \not\models C_P \sqsubseteq \top[0, 0]$. Eine Assertionalkomponente \mathcal{A} ist genau dann von \mathcal{T}_P erfüllt, wenn eine probabilistische Interpretation \mathcal{P}_I existiert, die Modell von \mathcal{T}_P ist.

Definition 3.16: probabilistische Subsumption Für eine Begriffsbeschreibung D , welche die Begriffsbeschreibung C in einer probabilistischen Terminologiekomponente \mathcal{T}_P subsumiert, wird der Grad der Überdeckungswahrscheinlichkeit im Intervall $[u, o]$ von D über C nach $\mathcal{T}_P \models (D|C)[u, o]$ berechnet. Ebenfalls wird für eine Begriffsbeschreibung C , dem bestimmten Individuum i_1 und der abstrakten Rolle $R \in \mathbf{R}_{\mathbf{A}}$ das Wahrscheinlichkeitsintervall $[u, o]$ nach $\mathcal{T}_P \models (\exists R.\{i_1\}|C)[u, o]$ berechnet, dass i_1 über R mit einem beliebigen Individuum von C in Beziehung steht.

Definition 3.17: probabilistische Instanz Für ein bestimmtes Individuum $i_1 \in \mathbf{I}$ und einer Begriffsbeschreibung C in einer probabilistischen Terminologiekomponente \mathcal{T}_P wird das Wahrscheinlichkeitsintervall $[u, o]$ nach $\mathcal{T}_P \models (C|\{i_1\})[u, o]$ berechnet, mit dem i_1 von C beschrieben wird. Ebenfalls wird für zwei Individuen i_1, i_2 und der abstrakten Rolle $R \in \mathbf{R}_{\mathbf{A}}$ das Wahrscheinlichkeitsintervall $[l, u]$ nach $\mathcal{T}_P \models (\exists R.\{i_2\}|\{i_1\})[u, o]$ berechnet, in dem i_1 über R mit i_2 in Beziehung steht.

3.3.2.3 Einbettung der Wahrscheinlichkeiten in OWL mittels Reifikation

Die Sprache OWL kennt einen vordefinierten Satz von Datentypen \mathbf{D} . Diese umfassen gängige Datentypen wie die Menge der reellen Zahlen oder der Zeichenketten. Zur Repräsentation

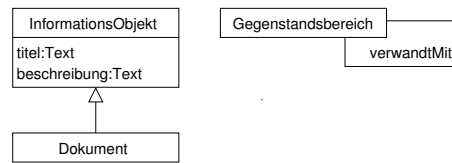


Abbildung 3.6: Minimal benötigte Abstraktionen

des Wahrscheinlichkeitsintervalls $[u, o]$ erfolgt die Spezifikation zweier generischer Attribute `untereProbabilitätsGrenze` und `obereProbabilitätsGrenze`. Diese können auf alle Elemente $o \in C \cap I \cap R$ einer Ontologie angewendet werden. Diese definieren sich zu:

Definition 3.18: Wahrscheinlichkeitsattribute Die Wahrscheinlichkeitsattribute `untereProbabilitätsGrenze` und `obereProbabilitätsGrenze` besitzen als Definitionsbereich den universellen Begriff \top und als Wertebereich den Datentyp der reellen Zahlen \mathbb{R} nach

- $\top \sqsubseteq \forall \text{untereProbabilitätsGrenze}.\top$ und $\exists \text{untereProbabilitätsGrenze} \sqsubseteq \mathbb{R}$.
- $\top \sqsubseteq \forall \text{obereProbabilitätsGrenze}.\top$ und $\exists \text{obereProbabilitätsGrenze} \sqsubseteq \mathbb{R}$.

Da sich sämtliche Relationen, wie auch die Subklassen- oder Instanzbeziehung, in OWL als Subjekt-Prädikat-Objekt Tripel eines RDF-Graphen manifestieren, lassen sich diese per Reifikation als Objekte behandeln (vgl. Abbildung 3.4, S. 76). Auf diese Weise lässt sich das Intervall $[u, o]$ für alle Subjekt-Prädikat-Objekt Tripel mit den oben definierten Attributen auszeichnen. Der Definitions- und Wertebereich schränkt dies auf Begriffsbeschreibungen und Individuen ein.

3.3.3 Minimalabstraktionen

Über die Spezifikationssprache pOWL lassen sich generische probabilistische Ontologien formulieren und unter Berücksichtigung der genannten Randbedingungen (vgl. 3.3.2.1, S. 82) ineinander einbinden. Da der Hauptanwendungsfall eines IR-Systems die Recherche von textueller Information ist, werden drei probabilistische Begriffe als Minimalstruktur eines Dokumentenraummodells vorgegeben. Ansonsten können die verwendeten Ontologien frei gestaltet sein. Die damit modellierten Inhalte sind damit vollständig zugänglich und auf eine Visualisierung abbildbar. Die Minimalabstraktionen beziehen sich als strukturelle Basis auf eine Klassifikation anhand des inhaltlich-strukturellen Typs eines Dokuments und seines Gegenstandsbereichs. Die minimalen Basiselemente `InformationsObjekt`, `Dokument`, `Gegenstandsbereich` und die transitive Rolle `verwandtMit` über den `Gegenstandsbereich` sind als UML-Klassendiagramm¹⁴ in Abbildung 3.6 dargestellt.

Der probabilistische Begriff `InformationsObjekt` bezeichnet in diesem Kontext die Menge aller Inhalte D , die im Umfeld einer Recherche von einem Nutzer zu einer Gesamtinformation konstruiert werden können. Er dient als Oberbegriff der vom Anwendungsfall abhängigen spezialisierten Typen von Informationsobjekten. Neben `Dokument` ist zum Beispiel der Begriff `Webseite` denkbar. Im Minimalfall besitzt ein `InformationsObjekt` zwei extrinsische und nicht probabilistische Attribute: einen Titel (`titel`) und eine kurze textuelle Beschreibung (`beschreibung`). Der Titel gibt den Namen eines `InformationsObjekt` an, unter dem es allgemein bekannt ist. In dem Attribut `beschreibung` ist eine knappe inhaltliche Zusammenfassung des Informationsobjekts hinterlegt.

Ein `InformationsObjekt` zerfällt zumindest in den Unterbegriff `Dokument`, welcher die Menge aller über eine Stichwortsuche auffindbaren Dokumente bezeichnet. Dieser erbt gemäß der modellierten Klassenhierarchie die beiden Attribute `titel` und `beschreibung`. Das Attribut `beschreibung` und der Begriff `Dokument`, werden hauptsächlich für die Funktionalitäten der Nutzerschnittstelle (vgl. Abschnitt 3.5, S. 98) benötigt.

¹⁴OBJECT MANAGEMENT GROUP: *The Unified Modeling Language (UML) Specification*, 2001

Im Weiteren existiert der probabilistische Begriff *Gegenstandsbereich*, welcher anhand der jeweiligen Diskurswelt Teilmengen der durch *InformationsObjekt* bezeichneten Individuen thematisch klassifiziert. Alle Themengebiete des *Gegenstandsbereichs* sind Subbegriffe von *Gegenstandsbereich*. Die transitive und probabilistische Rolle *verwandtMit* weist dem *Gegenstandsbereich* eine reflexive Verwandtschaftsbeziehung zu, mit der die klassifizierten Inhalte als "sich inhaltlich ähnlich" ausgezeichnet werden dürfen. Nicht nur Themen können damit in Beziehung gesetzt werden, sondern auch einzelne Informationsobjekte. Die Wahrscheinlichkeitsgewichte u, o der Beziehung sind hierbei als Verwandtschaftsgrad zu interpretieren. Diese Verwandtschaftsbeziehung kann ebenfalls als Superrolle all derjenigen Rollen aufgefasst werden, mit denen Beziehungen von Themengebieten näher spezifiziert werden. Bei Bedarf lassen sich problemlos auch weitere, von *verwandtMit* unabhängige Rollen zur Auszeichnung von Themengebieten einführen, die nicht notwendigerweise die Eigenschaft der Transitivität aufweisen.

Formal ergibt sich die folgende Minimalterminologie:

$$\begin{aligned} \mathcal{T}_{Min} = & \{(\text{InformationsObjekt} \doteq (\exists_{=1} \text{titel.Text}) \\ & \cap (\exists_{=1} \text{beschreibung.Text})), \\ & (\text{Dokument} \subseteq \text{InformationsObjekt}), \\ & (\exists \text{verwandtMit}^{S+} \subseteq \text{Gegenstandsbereich}), \\ & (\top \subseteq \text{Gegenstandsbereich}) \end{aligned} \quad (3.1)$$

Inhalte D sind damit mindestens doppelt klassifiziert. Im Optimalfall ist jedes Individuum von *InformationsObjekt* auch Teil von *Gegenstandsbereich*, so dass $\text{InformationsObjekt} \cap \text{Gegenstandsbereich} = \text{InformationsObjekt}$. Es sei darauf hingewiesen, dass dieses nicht zwingend vorgegeben ist. Gerade bei hochdynamischen Informationsräumen wie Intranets sind Texte meist zuerst über eine automatisierte Stichwortsuche recherchierbar, während eine vollständige thematische Klassifikation aufgrund einer intellektuellen Überarbeitung erst später nachgezogen wird. Dies bedeutet, dass für noch nicht thematisch zugeordnete Inhalte eine semantische Suche nur auf Grundlage der formalen Gegebenheiten stattfinden kann. Zu diesem Zeitpunkt lassen sich die komplexen semantischen Beziehungen des modellierten Informationsraums für die Recherche von solchen Dokumenten noch nicht ausnutzen.

Eine alternative Methode zur Modellierung von Informationsobjekten, welche in einer Themenhierarchie klassifiziert sind, wäre die Bildung von den Begriffen *InformationsObjekt* und *Thema*, wobei die einzelnen Themengebiete Individuen von *Thema* sind¹⁵. Mittels der transitiven und reflexiven Rolle *istUnterThemaVon* über *Thema* lässt sich auf der Individualebene aussagen, dass ein bestimmtes Thema Unterthema eines anderen ist. Auf diese Weise lässt sich ebenfalls eine Themenhierarchie bilden. Diese Modellierungspraxis ist sehr dokumentenzentriert. Eine reichhaltigere Auszeichnung von Querbeziehungen der einzelnen Themen und klassifizierten Inhalte untereinander lässt sich damit nicht erreichen. Während verwandte Themen sich in diesem Modell noch über die Rolle *verwandtMit* kennzeichnen ließen, wird damit nicht erlaubt dieselbe Relation auch auf die damit klassifizierten Informationsobjekte anzuwenden. Nun könnte die *verwandtMit*-Rolle auch auf den Begriff *InformationsObjekt* zugelassen werden. Damit wird jedoch nicht eindeutig klar, dass nur diejenigen Informationsobjekte, bei denen die zugehörigen Themengebiete in einer Verwandtschaftsbeziehung stehen, auch in einer Verwandtschaftsbeziehung stehen dürfen. Daher bietet sich die Spezifikation von Themengebiete und deren Zusammenhänge als eigenständige Begriffsbeschreibungen an.

Damit die Ontologien von Informationsobjekten einen möglichst hohen Grad an Standardisierung aufweisen, wird für die syntaktische Beschreibung von Rollen der Informationsobjekte der Metadatenstandard *Dublin Core*¹⁶ (DC) herangezogen. *Dublin Core* ist ein in ISO 15836-2003¹⁷ spezifizierter Standard zur anwendungsunabhängigen Auszeichnung aller Arten von Inhalten. Der Standard besteht aus 15 Elementen, welche zur Beschreibung der unterschiedli-

¹⁵MOTIK/MAEDCHE/VOLZ: *A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications*, 2002

¹⁶DUBLIN CORE METADATA INITIATIVE: *Dublin Core Metadata Element Set, Version 1.1: Reference Description*, 2004

¹⁷INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: *ISO 15836: Information and documentation - The Dublin Core metadata element set*, 2003

DC Element	Beschreibung	Verwendung als
dc:title	Eine allgemeine Bezeichnung eines Informationsobjekts.	Attribut von InformationsObjekt
dc:description	Eine inhaltliche Kurzbeschreibung.	Attribut von InformationsObjekt
dc:identifier	Ein eindeutiger Bezeichner innerhalb eines gegebenen Kontexts.	Relation zu Individuum
dc:creator	Ein Ersteller (Person, Organisation, etc.) des gegebenen Informationsobjekts.	Relation zu Individuum
dc:subject	Ein Bezeichner für den direkten Betreff eines Informationsobjekts.	Relation zu Individuum
dc:publisher	Ein Herausgeber des gegebenen Informationsobjekts, der es verfügbar macht.	Relation zu Individuum
dc:contributor	Ein bei der Erstellung des Informationsobjekts Mitwirkender.	Relation zu Individuum
dc:date	Ein Datum im Lebenszyklus des Informationsobjekts (Erstellung, Herausgabe).	Relation zu Individuum
dc:type	Der inhaltlich-strukturelle Typ des Informationsobjekts.	Subbegriff von InformationsObjekt
dc:format	Ein Bezeichner für das physische oder elektronische Format des Informationsobjekts.	Relation zu Individuum
dc:source	Eine Referenz auf ein Objekt von dem das gegebene Informationsobjekt abgeleitet wurde.	Relation zu Individuum
dc:language	Ein Bezeichner für die Sprache in dem ein Informationsobjekt vorliegt.	Relation zu Individuum
dc:relation	Eine Referenz auf ein verwandtes Informationsobjekt.	Relation zu Individuum
dc:coverage	Eine Beschreibung der räumlichen oder zeitlichen Abdeckung.	Relation zu Individuum
dc:rights	Eine Beschreibung der rechtlichen Sachlage (Urheber-, Verlagsrecht).	Relation zu Individuum

Tabelle 3.3: Menge der Dublin Core Elemente

chen Metadaten herangezogen werden können. Die Bezeichner dieser Elemente sind als so genannte URIs¹⁸ gegeben, weshalb sie sich besonders für die Verwendung innerhalb von XML-Repräsentationen eignen. Die URIs der Elemente können direkt als Bezeichner für die jeweiligen Rollen der Attribute und Begriffe herangezogen werden. Zusätzlich wird für alle Objekte der Ontologien lexikalische Information hinterlegt, so dass eine visuelle Repräsentation in der jeweiligen Anwendungssprache gegeben ist. In diesem Sinne ist zum Beispiel die `verwandtMit` Rolle als `dc:relation` in der Ontologie modelliert, während `verwandtMit` der deutsche lexikalische Bezeichner für dieselbe ist. Ähnliches gilt für die Attribute `titel` und `beschreibung`. Eine Übersicht der Dublin Core Elemente zeigt Tabelle 3.3, wobei das URI-Prefix "<http://purl.org/dc/elements/1.1/>" der Elemente mit `dc:` abgekürzt wurde.

Die eingeführten Minimalabstraktionen erlauben, zusammen mit dem Instrument der Ontologieeinbettung, reichhaltige Modellierungsmöglichkeiten für Informationsräume. Abbildung 3.7 zeigt ein Beispiel für den Anwendungsbereich Wissensdatenbank. Im Beispiel werden eine Unternehmensontologie \mathcal{O}_{Unt} , eine Applikationsontologie der Domäne der Antriebe \mathcal{O}_{App} , eine Dokumentenontologie \mathcal{O}_{Doc} und die Minimalterminologie \mathcal{T}_{Min} zu einer Gesamtontologie zusam-

¹⁸im Englischen *Uniform Resource Identifier*. Siehe auch RFC 1630 des W3C

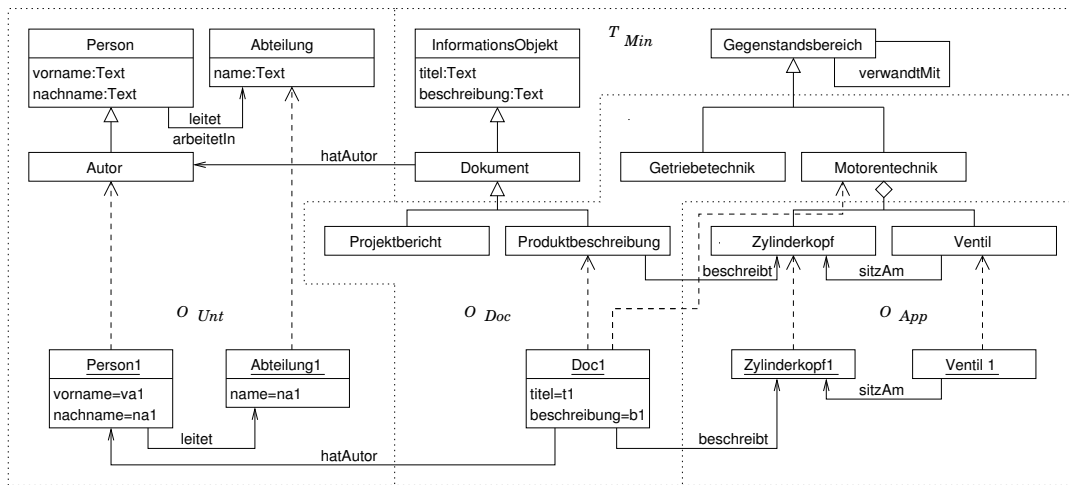


Abbildung 3.7: Modellierungsbeispiel einer Wissensdatenbank

mengeführt, welche zur Suche herangezogen wird. Dabei bindet die Dokumentenontologie die Minimalterminologie ein, während dieses Konstrukt von der Unternehmens- und der Applikationsontologie eingebunden wird.

Die Terminologie der Unternehmensontologie O_{Unt} definiert die Begriffe *Person*, *Abteilung* und *Autor*, wobei Autoren auch Personen sein müssen. Ferner erlauben die Rollen *arbeitetIn* und *leitet* die Auszeichnung von Personen, die in Abteilungen arbeiten beziehungsweise diese leiten. Das Individuum *Person1* leitet die *Abteilung1* und ist *Autor* von der *Produktbeschreibung doc1*. Die Applikationsontologie O_{App} sagt mittels der Begriffe *Ventil* und *Zylinderkopf*, sowie der Rolle *sitztAm* aus, dass *Ventil1* am entsprechenden *Zylinderkopf1* sitzt. Der Gegenstandsbereich der Dokumentenontologie O_{Doc} zerfällt in diesem Beispiel in *Getriebetechnik* und *Motorentechnik*, während *Produktbeschreibung* und *Projektberichte* Typen von *Dokument* sind. Über die Einbettung der jeweiligen Ontologien besitzt im Weiteren die *Produktbeschreibung Doc1* über die Rolle *hatAutor* den *Autor Person1* und *beschreibt* den speziellen *Zylinderkopf Zylinderkopf1*. Außerdem ist *Doc1* ein Individuum der Menge aller *Produktbeschreibungen* und der *Motorentechnik*. *Zylinderköpfe* und *Ventile* sind wiederum Teilbegriffe von *Motorentechnik* des *Gegenstandsbereichs*, mit dem sie über eine Aggregationsbeziehung zusammenhängen.

3.3.4 Integriertes Suchmodell

Das integrierte Suchmodell stellt im gewählten Ansatz eine Kombination der beschriebenen probabilistischen Deduktionskomponente mit einer herkömmlichen Stichwortsuche dar. Ein textueller Inhalt kann damit sowohl über Stichwörter recherchiert werden, die in den Texten auftauchen, als auch über semantische Merkmale, welche sich aus dem Inhalt und seiner Form ergeben. Die beiden Suchkomponenten sind derart gekoppelt, dass sich zu einzelnen Ergebnisdokumenten einer Stichwortsuche sämtliche semantischen Merkmale aus der Ontologie und andersherum abfragen lassen.

Für die Realisierung dieser Kopplung der an sich von einander unabhängigen Deduktions- und Stichwortsuchkomponenten müssen die in beiden Komponenten recherchierbaren Inhalte D gleichermaßen eindeutig identifizierbar sein. Daher wird eine Funktion

$$f_{ID} : d \rightarrow id | d \in D \quad (3.2)$$

eingeführt, welche jeden Informationsträger $d \in D$ auf eine eindeutige Kennung in Form einer URI abbildet. Diese URI wird zur eindeutigen Identifizierung eines Individuums der Ontologie,

welches einen tatsächlichen Inhalt repräsentiert, als auch eines Objekts des Dokumentenindex der Volltextkomponente herangezogen.

3.3.4.1 Stichwortsuche

Die Stichwortsuche gleicht eine Suchanfrage auf Grundlage einer Retrievalfunktion mit dazu passenden Ergebnissen ab (vgl. Abschnitt 2.4.1, S. 35). Die Suchanfrage stellt dabei eine Menge von Stichwörtern dar. Das tatsächlich eingesetzte IR-Modell ist von dem hier beschriebenen Ansatz unabhängig, muss jedoch eine partielle Übereinstimmung von Anfrage und Ergebnis ermöglichen. Als einzig notwendiges Kriterium muss die Retrieval-Funktion R auf eine Anfrage q eine Menge von Suchergebnissen d zurückliefern, die nach Relevanz sortiert sind. Die Signatur der Retrievalfunktion der Stichwortsuchkomponente ergibt sich zu

$$R : q' \times d' \Rightarrow [0, 1]$$

Die Relevanzen der einzelnen Ergebnisse liegen damit im Intervall $[0, 1]$.

3.3.4.2 Semantische Suche

Die semantische Suche erfolgt durch die Konstruktion einer Begriffsbeschreibung, wobei für deren einzelne Elemente freie Variablen verwendet werden können, die dann von der Deduktionskomponente an Objekte der Ontologie gebunden werden, welche die Begriffsbeschreibung erfüllen. So bindet der Ausdruck $\{x\} \sqsubseteq C$ alle Individuen von C an die freie Variable x . Die Begriffsbeschreibungen können dabei beliebige Komplexität aufweisen. Zum Beispiel findet der Ausdruck $\{x\} \sqsubseteq ((\text{Produktbeschreibung} \sqcap \text{Motorenteknik}) \sqcap (\exists \text{beschreibt.Zylinderkopf}))$ in der in Abbildung 3.7 gezeigten Ontologie alle Individuen x , die eine Produktbeschreibung darstellen, in Motorenteknik eingeordnet sind und Zylinderköpfe beschreiben. Die Ergebnismenge dieser Anfrage ist $\{Doc1\}$. Andersherum lassen sich über $\{Doc1\} \sqsubseteq x$ alle zugehörigen Klassen C von $Doc1$ herausfinden.

3.4 Ein kollaboratives Verfahren zur Wissensakquise

Vor der eigentlichen Nutzung des semantischen Modells zur wissensbasierten Suche muss eine Abstraktion der Anwendungsdomäne unter Zuhilfenahme der Minimalstruktur erfolgen. Im Abschnitt 2.5.1.4, S. 49 wurden bestehende Ansätze zur initialen Wissensakquise vorgestellt, die eine solche Modellierung unterstützen. Diese gehen von der manuellen Erstellung durch einen speziellen Modellierer aus und beziehen eine möglichst große Zahl potentieller Nutzer in den Ontologieentwicklungsprozess ein, um einen breiten Konsens der modellierten Inhalte und der verwendeten Begrifflichkeiten zu erreichen. Dieser Konsens unter den Nutzern ist wichtig, da beim letztendlichen Gebrauch des resultierenden wissensbasierten Systems die verwendete Terminologie leicht verständlich und intuitiv strukturiert sein muss.

Für die maschinelle Unterstützung des Modellierers wird das im selben Abschnitt aufgezeigte Rahmenwerk zur Extraktion von Ontologien aus natürlichsprachlichen Texten¹⁹ herangezogen. Verschiedenste Techniken zur Verarbeitung der natürlichen Sprache und des maschinellen Lernens sind darin integriert, welche die Identifikation von relevanten Begrifflichkeiten und deren Beziehungen untereinander unterstützen. Jedoch ist es in diesem Rahmenwerk nicht möglich, bedeutungsvolle Schlagwörter beziehungsweise Begriffe für ganze Themengebiete zu extrahieren, wenn diese gar nicht erst in den analysierten Texten auftauchen. Auch ist nach einer initialen Erstellung die gebildete Ontologie als ein evolutionäres Gebilde aufzufassen²⁰, welches sich stetig weiter zu entwickeln hat. Veränderungen der damit modellierten Anwendungsdomäne sind

¹⁹MAEDCHE/STAAB: *Mining Ontologies from Text*, 2000

²⁰FERNANDEZ-LOPEZ/GOMEZ-PEREZ/JURISTO: *METHONTOLOGY: From Ontological Art Towards Ontological Engineering*, 1997

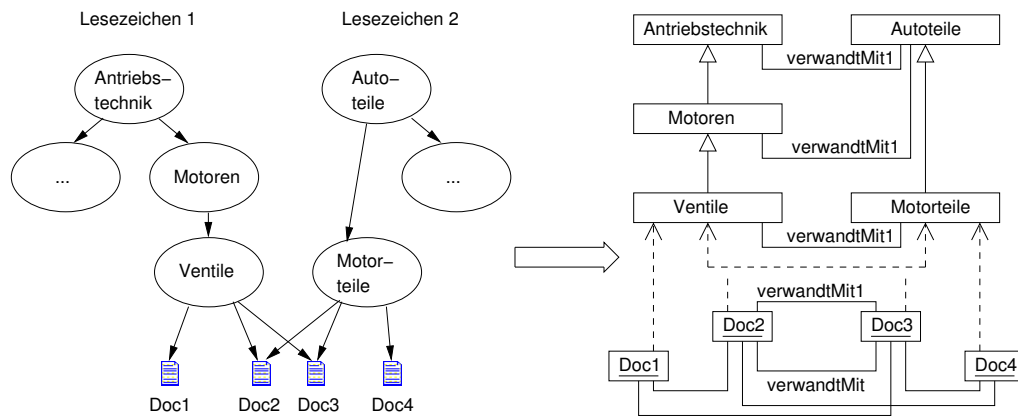


Abbildung 3.8: Grundlegende Idee der kollaborativen Wissensakquise

nachzuziehen. Gleichfalls gilt es, entsprechende Nutzerbedürfnisse zum Beispiel aufgrund von Verständnisproblemen aufzulösen und einzupflegen.

Das hier vorgestellte Verfahren des kollaborativen Indexierens unterstützt die initiale Erstellung und Pflege von Ontologien, in dem neues Systemwissen aus den Artefakten generiert wird, welche in den Kommunikationsprozessen zwischen den Nutzern und dem Rechtersystem entstehen. Diese Artefakte beziehen sich auf Informationsablagestrukturen einzelner Nutzer, in denen relevante Inhalte für diese Nutzer organisiert sind. Die Erzeugung von Ablagestrukturen seitens der Nutzer stellen Kontexte dar, in denen diese ihr Wissen über vorhandene Inhalte explizieren. Daher ist in der Nutzerschnittstelle eine spezielle Möglichkeit zur Ablage und Strukturierung von Lesezeichen vorgesehen.

Ablagestrukturen manifestieren sich als Lesezeichen, die in der Regel in entsprechend strukturierten Ordnern abgelegt sind. Sie dienen als Merkhilfe innerhalb eines Informationssystems. Ablagestrukturen können aber prinzipiell auch Verzeichnisstrukturen eines Dateisystems sein. Lesezeichen bestehen aus Referenzen auf bestehende Inhalte eines Informationsraums, welche sich zu einer besseren Wiederauffindbarkeit in Ordnern zusammenfassen lassen. Für die einzelnen Ordner können jeweils textuelle Bezeichner vergeben werden und sie lassen sich in einer Hierarchie strukturieren. Somit können Ablagestrukturen als persönliche Klassifikationssysteme betrachtet werden, welche für den jeweiligen Nutzer relevante Informationsobjekte beschreiben und kategorisieren. In diesen Ablagestrukturen wird das Nutzerwissen über die so klassifizierten Inhalte reflektiert. Diese Information ist im Kommunikationsprozess innerhalb eines soziotechnischen Systems vermittelbar und kann durch Interpretation für eine systemische Wissensbildung herangezogen werden.

Aus der Analyse solcher Klassifikationssysteme lassen sich daher mit einigen begründbaren Annahmen neue Aussagen über Inhalte und Begrifflichkeiten, sowie deren Ähnlichkeit ableiten. Die zugrunde liegende Idee zeigt Abbildung 3.8.

Innerhalb zweier Ablagestrukturen spiegelt sich ein Teilausschnitt der Domäne der Autos wieder. Die mit den Lesezeichen referenzierten Dokumente "Doc1", "Doc2", "Doc3" wurden unter dem Ordner "Ventile" und "Doc2", "Doc3", "Doc4" unter dem Ordner "Motorenteile" abgelegt. "Motorenteile" wurde dabei als Unterordner von "Autoteile" angelegt, während "Ventile" als Unterordner von "Motoren" und "Antriebstechnik" gewählt wurde. Aus der expliziten Kategorisierung von "Doc1", "Doc2", "Doc3" lässt sich sowohl auf eine Ähnlichkeit der Dokumente selbst als auch auf eine thematische Zuordnung zu dem Ordnungsbezeichner "Ventile" schließen. Gleiches gilt für den Ordner "Motorenteile" und den damit klassifizierten Dokumenten. Da die Ablagestrukturen in Ordner-Unterordner Hierarchien strukturiert wurden, können die übergeordneten Ordnungsbezeichner auch als thematische Oberbegriffe gewertet werden. Daraus ergibt sich wiederum eine thematische Zuordnung der einzelnen Dokumente zu den entsprechenden Oberbegriffen.

Die beiden Dokumente "Doc1" und "Doc2" wurden einmal unter dem Aspekt der "Motorenteile"

und ein anderes mal unter dem Aspekt der "Ventile" betrachtet. Neben der inhaltlichen Ähnlichkeit der beiden Dokumente kann daraus auf eine inhaltliche Verwandtschaft der beiden Begriffe "Motorteile" und "Ventile" geschlossen werden. Der Grad der Verwandtschaft kann anhand des Überschneidungsgrades, das heißt der Anzahl der gemeinsamen Lesezeichen, bestimmt werden. Diese Verwandtschaftsbeziehung propagiert sich mit zunehmender Abschwächung des Verwandtschaftsgrades in der Hierarchie nach oben, so dass ebenfalls ein geringer Zusammenhang zwischen Autoteile und Antriebstechnik gefolgert werden kann.

Auf ontologischer Seite ergeben sich aus den Ordnern einer Ablagestruktur Begriffsassertionen und atomare Begriffsbeschreibungen für die, mit den Lesezeichen korrespondierenden Individuen. Aus den Ordner-Unterdner Hierarchien werden Subsumptionsaxiome für diese Begriffsbeschreibungen. Die Überlagerung der Ablagestrukturen liefert entsprechende Querbeziehungen beziehungsweise entsprechende Rollen für die Begriffsbeschreibungen und für einzelne Individuen. Die bedingten Wahrscheinlichkeiten der Instanz-, Subsumptions- und Querbeziehungen zwischen Individuen und Begriffsbeschreibungen werden durch eine statistische Analyse aller zugänglichen Ablagestrukturen ermittelt. Damit ergibt sich die ebenfalls in Abbildung 3.8 gezeigte probabilistische Ontologie.

Bei den generierten Querbeziehungen sind zwei Arten zu unterscheiden: Zum einen ergibt sich eine inhaltliche Verwandtschaftsbeziehung der Dokumente selbst, welche über Lesezeichen in einem Ordner referenziert wurden. Der Beziehungstyp entspricht dabei der generischen *verwandtMit*-Rolle der Minimalabstraktionen (vgl. Abschnitt 3.3.3). Zum anderen werden Begriffsbeschreibungen zueinander mit neuen Beziehungstypen in Beziehung gesetzt. Diese deuten zwar auf eine Verwandtschaft der Begriffsbeschreibungen hin, sind aber semantisch aus der Analyse nicht eindeutig zu typisieren. Daher werden als Vorgabe zwischen zwei Begriffsbeschreibungen und den gemeinsamen Individuen jeweils neue Subrollen $\text{verwandtMit}_X \sqsubseteq \text{verwandtMit}$ der generischen Verwandtschaftsbeziehung eingezogen. Der Index X kennzeichnet dabei eine fortlaufende Nummer. Eine spätere genaue semantische Typisierung der betreffenden Rollen kann dann von einem menschlichen Autor vorgenommen werden.

Mit diesem Standardvorgehen der Vergabe von Rollen wird die semantische Konsistenz der Wissensbasis ohne Zuhilfenahme eines Autors sichergestellt. So werden im obigen Beispiel die beiden Dokumente *Doc2* und *Doc3* in eine Suche nach Dokumenten über *Ventile*, die mit *Motorteile* verwandt sind, einbezogen. Aber auch *Doc1* und *Doc4* werden erkannt, da eine Beziehung vom Typ *verwandtMit1* gefolgert werden kann. Die abgeleiteten Beziehungswahrscheinlichkeiten ergeben sich dabei durch Aufmultiplizieren der bedingten Wahrscheinlichkeiten der Beziehungen des Typs *verwandtMit1* und *verwandtMit*.

Die einzelnen Verwandtschaftsbeziehungen und -typen der Dokumente untereinander können so direkt zu Suche verwendet werden. Zusätzlich bieten sie eine Hilfe bei einer späteren intellektuellen Überarbeitung der so entstandenen Ontologien durch einen menschlichen Autor. Es können nicht nur die tatsächlichen semantischen Bezüge der Begriffe typisiert werden, auch Hinweise lassen sich liefern, für welche Individuen entsprechende Begriffsbeschreibungen gelten.

3.4.1 Grundlegende Hypothesen und Argumentation

In dem Verfahren des kollaborativen Indexierens werden Informationsablagestrukturen als persönliche Klassifikationssysteme der jeweiligen Nutzer betrachtet, aus denen ontologisches Wissen abgeleitet werden kann. In einer Ablagestruktur sind jedoch nicht notwendigerweise alle Lesezeichen in Ordnern strukturiert. So können Lesezeichen auch ohne Zuhilfenahme von Ordnern als flache Liste abgelegt sein, wie es zum Beispiel in Internetbrowsern möglich ist.

Auch entsteht Unschärfe aus der freien Strukturierungsmöglichkeit der Ordnerhierarchie und aus der Begriffsvergabe an diese Ordner, wenn Ober- und Unterbegriffe vertauscht werden oder die Ordnungsbezeichner umgangssprachlich beziehungsweise als Abkürzung gewählt werden.

Daher werden folgende Hypothesen bei der Verarbeitung von Ablagestrukturen aufgestellt:

Hypothese 1: Wird ein Lesezeichen für ein Informationsobjekt erstellt, so besitzt dieses Informationsobjekt für den jeweiligen Nutzer eine gewisse Wichtigkeit.

Hypothese 2: In einem benannten Ordner finden sich nur inhaltlich zusammenhängende Informationsobjekte.

Hypothese 3: Die Bezeichnung eines Ordners stellt einen gemeinsamen thematischen Oberbegriff der darunter organisierten Informationsobjekte dar.

Hypothese 4: Finden sich gleiche Informationsobjekte in verschiedenen Ordnern, bedeutet dies eine inhaltliche Verwandtschaft der Ordner.

Hypothese 5: Innerhalb der Menge der ohne Ordner strukturierten Lesezeichen finden sich inhaltlich zusammenhängende Informationsobjekte.

Hypothese 6: Gleiche Ordnungsbezeichner stellen Synonyme dar.

Hypothese 7: Für die überwiegende Mehrzahl der Ordnungsbezeichner werden allgemein sinnfällige thematische Begriffe vergeben.

Hypothese 8: Die überwiegende Mehrzahl der Ober-Unternordner Hierarchien ist so organisiert, dass der Bezeichner des Oberordners allgemeiner ist als der des Unternordners.

Die Extraktion von systemischem Wissen geschieht unter zwei Grundannahmen: Zum einen bringt eine gewisse Anzahl der Nutzer ihre Ablagestruktur in eine hierarchische Form, welche eine Klassifikation von Inhalten darstellt und damit das Auswerten von probabilistischen Begriffsbeschreibungen ermöglicht. Zum anderen stellen Lesezeichen in Ablagestrukturen und eventuell vorhandene Ordner Indikatoren für die Relevanz und den inhaltlichen Zusammenhang der abgelegten Informationsobjekte dar.

Die erste dieser Grundannahmen kann mit einer Studie von Boardman²¹ belegt werden, welche den Gebrauch von Hierarchien am elektronischen Arbeitsplatz untersuchte. Gemäß dieser Untersuchung bringen 50% der als fortgeschritten eingestufteten Nutzer ihre Arbeitsplatzressourcen wie Dateisystem, E-Post Programme und Browserlesezeichen in eine hochstrukturierte Form mit mehreren Hierarchieebenen. Weitere 30% der Nutzer legen immerhin eine Struktur mit einer Hierarchieebene an, während nur 20% der Nutzer ihre Ressourcen unstrukturiert ablegen und sie durch räumliche Zuordnung wieder finden.

Die zweite Annahme des inhaltlichen Zusammenhangs der in Ablagestrukturen organisierten Lesezeichen läßt sich dadurch argumentieren, dass Lesezeichen Resultat einer sehr intentionalen Handlung sind, welche einen gewissen Grad an Reflexionsleistung bedingen. Dies gilt besonders in dem Fall, wenn speziell benannte Ordner für die Organisation von Lesezeichen angelegt werden. Aber auch nicht klassifizierte Lesezeichen stellen einen Akt positivistischer Handlung der Interessensbekundung dar.

Das Erzeugen von benannten Ordnern und Ordnerhierarchien kann in etwa mit einer unmoderierten Mischung aus den Techniken der freien Auflistung²² (Free Listing) und des Kartensortierens²³ (Card Sorting) verglichen werden. Diese beiden empirisch wohlbelegten Methoden werden gerne beim Erstellen von Informationsarchitekturen, das heißt bei der begrifflichen Strukturierung von bestimmten Anwendungsbereichen für Informationssysteme im Umfeld der Mensch-Technik Interaktion eingesetzt. Die Methode der freien Auflistung versucht die inhaltliche Bedeutung und Abgrenzung einer Anwendungsdomäne zu erfassen. Dazu werden potentielle Nutzer eines Informationssystem aufgefordert, für bestimmte Ausschnitte des Anwendungsbereichs frei assoziierte Wörter aufzuschreiben. Auf der Grundlage öfters auftauchender Begriffe können diese als relevante Beschreibungen herangezogen werden.

In Ergänzung dazu kann mit der Technik des Sortierens von Karten herausgefunden werden, wie die jeweiligen Begriffe strukturiert werden. Meist wird diese Strukturierung für die Navigation in einem Informationssystem herangezogen. Ausgangspunkt dabei sind Kärtchen mit vordefinierten Begriffen. Bei der Durchführung werden die Begriffe gemischt und gut sichtbar vor einem Probanden aufgelegt. Dieser wird beauftragt, die Begriffe auf einer für ihn sinnfälligen Weise zu

²¹ BOARDMAN: *Multiple Hierachies in User Workspace*, 2001

²² BORGATTI: *Elicitation techniques for cultural domain analysis*, 1998

²³ ROSENFELD/MORVILLE: *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites*, 2002

sortieren. Zusammen gehörende Begriffe werden jeweils auf einen eigenen Stapel gelegt und stellen eine eigene Kategorie dar. Jede Gruppierung wird vom Probanden mit einem Überbegriff benannt.

Auf dieser Grundlage besteht die begründete Annahme, gravierende Unschärfen sowohl bei der Begriffswahl als auch bei der hierarchischen Strukturierung statistisch heraus filtern zu können, wenn eine genügend hohe Anzahl von Ablagestrukturen betrachtet wird. Ebenfalls werden Hierarchien anhand der Häufigkeit aufgebaut, mit der Ordnungsbezeichner übereinander angeordnet werden. Begriffsbeschreibungen und Rollen, welche aufgrund ihrer Häufigkeitsverteilungen nicht signifikant heraus treten, werden aus dem Ontologiegraphen wieder entfernt.

3.4.2 Aufgabenstellung

Wie zuvor erwähnt, werden bei dieser Überlagerung die textuellen Bezeichner der jeweiligen Ordner in Begriffsbeschreibungen für die damit indexierten Lesezeichen überführt. Die Lesezeichen stellen dann die Individuen der jeweiligen Begriffsbeschreibungen dar. Für alle Lesezeichen innerhalb eines Ordners wird auf eine Verwandtschaftsbeziehung der korrespondierenden Individuen geschlossen. Gleichfalls wird auf eine Verwandtschaftsbeziehung von Begriffsbeschreibungen geschlossen, wenn verschiedene Ordner gemeinsame Lesezeichen enthalten und sich überschneiden. Aus der expliziten hierarchischen Strukturierung von Ordnern wird eine Subsumptionsbeziehung der korrespondierenden Begriffsbeschreibungen gefolgert. Die bedingten Wahrscheinlichkeiten der jeweiligen Beziehungstypen werden durch auszählen bestimmt. Sie stellen damit ein kollaboratives Ähnlichkeitsmaß dar.

Die grundsätzliche Aufgabenstellung besteht darin, die generischen probabilistischen Randbedingungen $[u, l]$ von Begriffsbeschreibungen einer in pOWL ausgedrückten Ontologie zu bestimmen. Da jedoch kein Wahrscheinlichkeitsintervall bestimmt werden kann, sondern nur ein Wahrscheinlichkeitswert, werden im Folgenden die probabilistischen Randbedingungen $[u, l]$ zu $[p, p]$ beziehungsweise zu $[p]$ abgekürzt. Im Detail gilt es daher, die Wahrscheinlichkeitswerte $[p]$ zu bestimmen, mit der

- $(D|C)[p]$ die Begriffsbeschreibung D die Begriffsbeschreibung C subsumiert.
- $(C|i_1)[p]$ das Individuum i_1 tatsächlich Individuum der Begriffsklasse C ist.
- $\exists R.\{C\}\{D\}[p]$ Individuen der Begriffsklasse C existieren, die mit der Rolle R mit Individuen der Begriffsklasse D in Beziehung stehen.
- $\exists R.\{i_1\}\{i_2\}[p]$ ein bestimmtes Individuum i_1 mit der Rolle R mit dem bestimmten Individuum i_2 in Beziehung steht.

Ferner sind Häufigkeitswerte für die entstandenen Begriffsbeschreibungen und Individuen abzuleiten, damit eine Filterung von Unschärfen erfolgen kann. Diese Häufigkeitswerte sind als Relevanzwerte zu interpretieren, da sie nicht in eine, in der Ontologie gültigen Interpretation von Wahrscheinlichkeit überführt werden können.

3.4.3 Kollaboratives Indexieren

Die hier betrachteten Ablagestrukturen bestehen prinzipiell aus einer Menge von Lesezeichen, welche als Merkhilfe innerhalb eines Informationssystems angelegt wurden. Diese Menge der Lesezeichen kann, muss aber nicht zwingend, mit Hilfe von bezeichneten Ordnern hierarchisch strukturiert sein. Auf oberster Ebene einer Ablagestruktur können sich als flache Liste abgelegte Lesezeichen befinden. Die Struktur von Lesezeichen lässt sich so auch auf Ordnerstrukturen eines Dateisystems übertragen.

Während die Bezeichner von Ordnern direkte Hinweise auf eine zu verwendende Begrifflichkeit liefern, fehlt für die als flache Liste abgelegten Lesezeichen eine solche Information. Die Behandlung von losen und in Ordnern organisierten Lesezeichen erfolgt deshalb mittels zweier unterschiedlicher Verfahren, um aus beiden Ablageformen systemisches Wissen zu extrahieren.

Für in Ordnern strukturierten Lesezeichen können die bestimmten Wahrscheinlichkeiten unter Zuhilfenahme der obigen Hypothesen 1, 2 und 3 durch Auszählen ermittelt werden.

Für die Entdeckung ähnlicher Informationsobjekte in flachen Listen existieren bereits effiziente Algorithmen. Diese entstammen der Warenkorbanalyse, einem klassischen Problem des Data Minings. Dabei werden Gesetzmäßigkeiten im Kaufverhalten von Kunden untersucht, um durch geeignete Aufstellung von Waren das Umsatzvolumen zu steigern. Aus diesen werden so genannte Assoziationsregeln²⁴ im Sinne von "wer Kaviar isst, trinkt dazu gerne Wodka" generiert.

Die aus beiden Verfahrenen gewonnenen bedingten Wahrscheinlichkeiten werden in der Ontologie zusammen geführt. Nach der formalen Definition einer Ablagestruktur erfolgt im Weiteren eine Darstellung der beiden Methoden.

3.4.3.1 Ablagestrukturen

Eine Ablagestruktur enthält auf oberster Ebene eine Menge von ungeordneten Elementen (Lesezeichen), sowie einer Menge von in Ordnern organisierten Elementen. Jedem Ordner ist ein textueller Bezeichner zugeordnet. Die jeweiligen Ordner können weitere Unterordner enthalten, die wiederum weitere Elemente enthalten.

Definition 3.19: Ablagestruktur: Eine Ablagestruktur $S \in \mathcal{S}$ aus der Menge aller Ablagestrukturen \mathcal{S} ist ein Paar $S = (E_0, \mathcal{F})$ wobei

- E_0 die endliche und möglicherweise leere Menge der ungeordneten Elemente und
- \mathcal{F} ein Wald von Ordnern darstellt.

Ein Ordner stellt sich mathematisch als Baum dar, wobei die Blätter des Baumes die damit geordneten Elemente oder leere Unterordner darstellen. Die Äste des Baumes können jedoch nur Unterordner sein. Jeder Unterordner stellt wiederum einen Baum dar, wenn er für sich alleine betrachtet wird. Die oberste Hierarchieebene der Ablagestruktur wird mit dem Index 0 gekennzeichnet. Ferner existiert eine Halbordnung der Ordner (Hierarchie) die mit dem Symbol \succ ausgedrückt wird.

Definition 3.20: Ordner Jeder Ordner $F \in \mathcal{F}$ ist ein Baum der Höhe \mathbb{H} mit der Wurzel $F_{0,0}$. Jede Ebene des Baumes hat die Breite $\mathbb{B}(\mathbb{H})$. Jeder Ast $F_{h,b}$ ist Teilbaum von F und stellt einen Ordnungsbezeichner dar, der eine Menge von Elementen $E_{h,b}$ ordnet. Die einzelnen Elemente $e_{h,b,i} \in E_{F_{h,b}}$ des Astes $F_{h,b}$ können nur dessen Blätter sein. Ferner existiert eine reflexive, antisymmetrische und transitive Relation über F , so dass sich eine Halbordnung auf $F = (F, \succ)$ ergibt.

3.4.3.2 Behandlung von geordneten Lesezeichen

Für die Behandlung eines Ordners $F_{h,b} \in \mathcal{F}$ einer Ablagestruktur S wird zunächst das jeweilige Auftreten der Ordnungsbezeichner und seiner Elemente in korrespondierenden Zählern $C_{F_{h,b}}(F_{h,b})$ und $C_E(e_{h,b,i})$ mitgezählt, wobei in einer weiteren Variablen $C_{F_{max}} = \max(C(F_{h,b}))$ der maximale Zählerstand mitgeführt wird. Der Relevanzwert für Ordnungsbezeichner wird durch

$$R(F_{h,b}) = \frac{C(F_{h,b})}{C_{F_{max}}} \quad (3.3)$$

berechnet. Diese Formel entspricht der normierten Termfrequenz des Information Retrievals (vgl. Abschnitt 2.4.2, S. 36).

Für alle Unterordner $F_{h+1,b}$ eines Ordners $F_{h,b}$ wird das gemeinsame Auftreten von Unter- und Oberordner in jeweils einem Zähler mitgezählt. Diese Zähler $C_{F_{j,b} \wedge F_{j-1}} = F_{j,b} \wedge F_{j-1, b}^{\bigwedge_{j=h}^{j=1}}$ werden für alle $F_{j,b}^{\bigwedge_{j=h}^{j=1}}$ mit $b_{\mathbb{B}(j)}^0$ berechnet. Dies bedeutet, dass nicht nur das gemeinsame

²⁴ AGRAWAL/IMIELINSKI/SWAMI: *Mining Association Rules between Sets of Items in large Databases*, 1993

Auftreten von einem Ordner zu seinem direkten Oberordnern gezählt wird, sondern zu allen seinen Oberordnern.

Die bedingte Wahrscheinlichkeit, dass ein Ordner $F_{h,b}$ einen anderen Ordner $F_{h+1,b}$ subsumiert, berechnet sich nach

$$P(F_{h,b}|F_{h+1,b}) = \frac{C_{F \wedge (F_{h,b} \wedge F_{h+1,b})}}{C_F(F_{h+1,b})} \quad (3.4)$$

Ein ähnliches Vorgehen wird für die einzelnen Elemente in einem Ordner angewandt. Nach Hypothese 2 wird für alle Elemente in einem Ordner deren gemeinsames Auftreten mitgezählt. Für zwei Elemente aus $e_{k,l} \in E_{h,b}$ existiert der Zähler $C_{E_{k \wedge l}} = e_k \wedge e_l \big|_{k,l=0}^{\|E_{h,b}\|}$. Die bedingten Wahrscheinlichkeiten der Beziehung verwandtMit für die beiden Elemente $e_{k,l} \in E_{h,b}$ berechnen sich zu

$$\begin{aligned} P(e_k|e_l) &= \frac{C_{E \wedge (e_k \wedge e_l)}}{C_E(e_l)} \\ P(e_l|e_k) &= \frac{C_{E \wedge (e_k \wedge e_l)}}{C_E(e_k)} \end{aligned} \quad (3.5)$$

Auf Grundlagen von Hypothese 3 wird zur Berechnung der jeweiligen Instanzbeziehungs-gewichte der einzelnen Elemente in einem Ordner das jeweils paarweise Auftreten von einem Element und seinem Ordner mitgezählt. Für ein Elemente aus $e_k \in E_{h,b}$ und einem Ordner $F_{h,b} \in F$ existiert der Zähler $C_{E \wedge F} = e_k \wedge F_{h,b} \big|_{k=0}^{\|E_{h,b}\|}$. Die bedingten Wahrscheinlichkeiten der Instanzbeziehung für die beiden Elemente $e_k \in E_{h,b}$ und $F_{h,b} \in F$ berechnen sich zu

$$P(e_k|F_{h,b}) = \frac{C_{E \wedge F}(e_k \wedge F_{h,b})}{C_E(e_k)} \quad (3.6)$$

Nach Hypothese 4 wird auf inhaltliche Verwandtschaft geschlossen, wenn sich Ordner durch gemeinsame Lesezeichen überschneiden. Wenn E_{F_X} die Menge der Elemente eines Ordners F_X und E_{F_Y} die Menge der Elemente eines Ordners F_Y darstellen, dann berechnen sich die Wahrscheinlichkeiten der jeweiligen inversen Verwandtschaftsbeziehungen verwandtMitXY und verwandtMitYX zu

$$\begin{aligned} P(F_Y|F_X) &= \frac{|E_{F_X} \cap E_{F_Y}|}{|E_{F_X}|} \\ P(F_X|F_Y) &= \frac{|E_{F_X} \cap E_{F_Y}|}{|E_{F_Y}|} \end{aligned} \quad (3.7)$$

Im ersten Durchlauf des Algorithmus werden die Relevanzmaße für die Ordnerbezeichner und die obigen Wahrscheinlichkeiten für alle Elemente der Lesezeichen S berechnet und in die Ontologie eingefügt. Im zweiten Durchlauf werden unter Zuhilfenahme der Relevanzmaße diejenigen Begriffsbeschreibungen herausgefiltert, deren Relevanzmaß unter einem vordefinierten Schwellwert R_{Min} liegt. Die Menge aller herauszufilternden Begriffsbeschreibungen C^F ergibt sich dann zu $C^F \subseteq C|R(F) < R_{Min}$. Die jeweiligen Individuen einer Begriffsbeschreibung $\{i_k\} \subseteq C_h^F[p]$ werden dabei der nächsten, in der Hierarchie weiter oben stehenden und nicht herausgefilterten Begriffsbeschreibungen zugeordnet, so dass das Individuum i_k zu einem Individuum von $\{i_k\} \subseteq C_{h-w}[p]$ wird. Die bedingten Wahrscheinlichkeiten der neuen Instanzbeziehung werden dabei anhand der Zähler nach Gleichung 3.6 neu berechnet.

Abbildung 3.9 verdeutlicht exemplarisch dieses Vorgehen in dem zwei Ablagestrukturen überlagert werden. Die beiden Ordner "Antriebstechnik" und "Motoren" tauchen dabei in beiden Ablagestrukturen auf. Der Zähler $C_{F_{max}}$ ist daher 2. Die Zählerstände der jeweiligen Ordnungsbezeichner und Beziehungen ergeben sich mit den in der Graphik dargestellten Werten. Die Re-

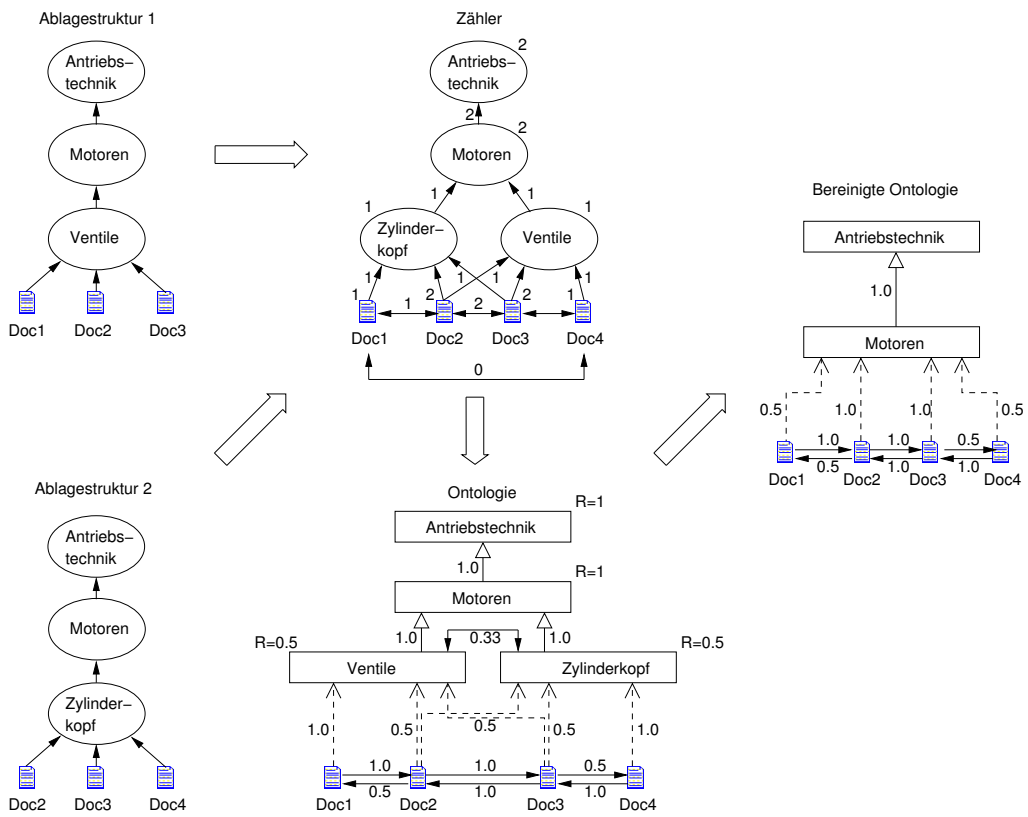


Abbildung 3.9: Beispielhafter Ablauf des kollaborativen Indexierens

levanzwerte der Ordnerbezeichner beziehungsweise der Begriffsbeschreibungen `Ventile` und `Zylinderkopf` ergeben sich zu 0.5. Bei einer Schwelle $R_{min} = 0.6$ werden diese beiden Begriffsbeschreibungen herausgefiltert. Interessant dabei ist, dass nach der Bereinigung der Ontologie die bedingten Wahrscheinlichkeiten der Instanzbeziehung von `Doc1` zu `Motoren` mit dem obigen Vorgehen auf 0.5 abschwächen.

3.4.3.3 Integration lose abgelegter Lesezeichen in die Ontologie

Nach Hypothese 5 existieren auch in der Menge nicht ungeordneten Lesezeichen \mathbf{E}_0 inhaltlich zusammenhängende Elemente, welche sich aufgrund eines häufigen gemeinsamen Auftretens bekannt machen. Das Entdecken dieser Elemente ist ein klassisches Problem des Data Mining, der Analyse von Warenkörben, und lässt sich mit dessen Methoden lösen. Die Menge \mathbf{E}_0 wird als Warenkorb beziehungsweise Transaktion interpretiert. Das Setzen eines Lesezeichens entspricht dann dem Kauf eines (Informations-) Produkts²⁵. Innerhalb der Menge aller Transaktionen $\mathbf{E}_0 \subseteq \mathcal{S}$ werden Assoziationsregeln zwischen zwei Elementen $e_1, e_2 \subseteq \mathbf{E}_0$ generiert. Die bedingten Wahrscheinlichkeiten für das Auftreten der beiden Elemente innerhalb einer Assoziationsregel wird auch Konfidenzfaktor der Assoziationsregel genannt.

Definition 3.21: Assoziationsregel Eine Assoziationsregel ist eine Implikation der Form $e_1 \rightarrow e_2$ mit der bedingten Wahrscheinlichkeit $P(e_1|e_2) = \frac{C_E(e_1 \wedge e_2)}{C_E(e_2)}$.

Da sich solche Assoziationsregeln für alle Elemente aus \mathbf{E}_0 generieren lassen, besteht das Problem darin, solche herauszufinden, die eine entsprechende statistische Signifikanz aufweisen. Daher wird der so genannte Unterstützungsfaktor eingeführt, der diese Signifikanz beschreibt.

Definition 3.22: Unterstützungsfaktor Der Unterstützungsfaktor einer Assoziationsregel ist der Quotient $\text{sup}(e_1 \rightarrow e_2) = \frac{C_E(e_1 \wedge e_2)}{|\mathbf{E}_0|}$ aus der Anzahl des gemeinsamen Auftretens von e_1 und e_2 mit der Gesamtzahl $|\mathbf{E}_0|$ aller Ablagestrukturen die eine nichtleere Menge \mathbf{E}_0 besitzen.

Der in der Literatur wohlbekannt Algorithmus APRIORI²⁶ berechnet nur solche Assoziationsregeln, deren Unterstützungsfaktor $\text{sup}(e_1 \rightarrow e_2) > \text{sup}_{min}$ größer als ein vorgegebener minimaler Unterstützungsfaktor sup_{min} ist.

Für alle Lesezeichen $e_i \subseteq \mathbf{E}_0$ werden solche Assoziationsregeln generiert. Eine Beziehung $\exists R.\{i_1\}|\{i_2\}[p]$ wird in die Ontologie aber nur dann eingefügt, wenn eine solche Beziehung nicht durch die Analyse geordneter Lesezeichen hergestellt werden konnte. Dies begründet sich dadurch, dass eine inhaltliche Ähnlichkeit von Lesezeichen eines Ordners der Ähnlichkeit aus einer Assoziationsregel vorzuziehen ist (Hypothese 2).

3.5 Schnittstelle zur Informationssuche

Die Nutzerschnittstelle ist die zentrale Komponente des, in Abschnitt 3.2, S. 69 vorgestellten Ansatzes, da mit deren Hilfe die Kommunikation zwischen Nutzer und Recherchesystem ermöglicht wird. In diesem Kommunikationsprozess können die von dem System verwalteten Inhalte als tatsächliche Information an den Nutzer vermittelt werden. Umgekehrt gestattet die Schnittstelle von der Systemseite aus die Beobachtung des Nutzers bei seinem Umgang mit recherchierter Information beziehungsweise deren Organisation. Beiden Kommunikationspartnern ist es möglich, aus der gewonnenen Information Wissen zu erzeugen.

Ferner sind innerhalb der Nutzerschnittstelle die grundsätzlichen Anforderungen an den gesamten Ansatz realisiert und der Zugriff auf die verschiedenen Systemfunktionalitäten integriert. Diese Anforderungen, die in dem vorhergehenden Kapitel beschrieben wurden, beziehen sich dabei auf die Bereitstellung von effektiven Zugangs- und Abfragemechanismen, die Förderung des

²⁵ GEYER-SCHULZ/HAHLER: *Evaluation of Recommender Algorithms for an Internet Information Broker based on Simple Association Rules and on the Repeat-Buying Theory*, 2002

²⁶ AGRAWAL/MIELINSKI/SWAMI: *Mining Association Rules between Sets of Items in large Databases*, 1993

Verständnisses des recherchierbaren Informationsraums und des eigenen Informationsbedarfs, sowie auf Hilfestellungen bei der inhaltlichen Erfassung der Ergebnismenge einer Suchanfrage.

Die Abfragemechanismen sind durch die Kopplung einer herkömmlichen Volltextsuchmaschine mit einer probabilistischen Deduktionskomponente realisiert. Die Volltextsuche ermöglicht eine ungenaue Suche auf Basis von Stichwörtern, während sich Inhalte mittels semantischer Abfragen anhand von bedeutungsvollen Merkmalen wie den Themenbezug oder struktureller Merkmale lokalisieren lassen. Die Stichwortsuchmaschine ist dabei mit der Deduktionskomponente derart verbunden, dass sich zu einzelnen Suchergebnissen sämtliche semantischen Merkmale aus der Ontologie abfragen lassen.

Zur Unterstützung eines besseren Verständnisses des Informationsraums wurde eine interaktive Visualisierung als Explorationsmöglichkeit von dessen semantischer Struktur gewählt. Als Grundlage der Strukturvisualisierung dient die ontologische Wissensbasis des Recherchesystems. Diese wird damit nicht nur zur logischen Abfrage herangezogen, sondern erfüllt gleichzeitig eine Transformations- und Verdichtungsfunktion bezüglich der verfügbaren Inhalte zur Visualisierung. Da die Strukturvisualisierung ein direktes Abbild der Wissensbasis darstellt, wird auf deren Grundlage die visuelle Konstruktion semantischer Abfragen ermöglicht. In der Folge entfällt damit die Notwendigkeit des Erlernens der eingesetzten logischen Abfragesprache. Durch die gleichzeitige Darstellung von Suchergebnissen und der semantischen Struktur erfolgt ferner eine Kontextualisierung von Suchergebnissen in dieser Struktur.

Die enge Verzahnung der einzelnen Visualisierungskomponenten schafft eine Vielzahl von ineinander greifender Interaktions-, Explorations- und Abfragemöglichkeiten. Unter deren Zuhilfenahme kann im Dialog und durch Rückkopplung mit den Antworten des Suchsystems der eigene, objektive Informationsbedarf ermittelt werden. Nur wenn der objektive Informationsbedarf des Nutzers erkannt wird, lassen sich die Anfragen an das System so formulieren, dass eine Informationsversorgung stattfinden kann. Die einzelnen Visualisierungselemente sind dabei in einer einheitlichen graphischen Nutzerschnittstelle integriert, um mit möglichst wenig Kontextwechseln auszukommen und damit den Nutzer möglichst wenig mental zu belasten.

Der tatsächliche graphische und interaktive Entwurf der Nutzerschnittstelle erfolgt unter dem Paradigma der konsequenten Verwendung allgemein bekannter und etablierter Metaphern und Interaktionstechniken. Damit wird der Lernaufwand im Umgang mit einer Implementierung des Recherchesystems verringert. Im Folgenden werden die eingesetzten Techniken beschrieben.

3.5.1 Visualisierung und Navigation ontologischer Information - der MatrixBrowser

Grundsätzlich lässt sich eine Ontologie als vernetzte Begriffsstruktur durch einen Graph darstellen. Dabei sind die einzelnen Begriffe die Knoten des Graphen, während die semantisch ausgezeichneten Relationen die Kanten des Graphen repräsentieren. Die Grundprinzipien einiger häufig gebrauchter Visualisierungstechniken zur Darstellung von vernetzten Begriffssystemen, wie sie durch Ontologien formalisiert werden, wurden im Abschnitt 2.6.1, S. 59 vorgestellt. Keiner der darin beschriebenen Ansätze zur Graphvisualisierung unterstützt in ausreichender Form die Darstellung beliebig vernetzter Informationsstrukturen im Hinblick auf unterschiedliche Nutzeraufgaben, wie die Suche nach spezifischen Knoten und Kanten, das Entdecken der Beziehungen zwischen beliebigen Knoten, oder die Erfassung aller Beziehungen eines bestimmten Knotens. Weiterhin fehlen wesentliche interaktive Eigenschaften, wie ein systematisches Vordringen in beziehungsweise die Verdichtung von Teilstrukturen.

Da aus diesen Gründen die existierenden Techniken nicht ausreichend geeignet sind, um die komplexen Netzstrukturen von Ontologien zu visualisieren und zu explorieren²⁷, erfolgt die Visualisierung der semantischen Struktur des Informationsraums beziehungsweise der ontologischen Wissensbasis mit Hilfe einer neuartigen Visualisierungstechnik, dem MatrixBrowser. Dieser ist als eine alternative Form der Graphendarstellung und -exploration zu verstehen.

²⁷ZIEGLER/KUNZ/BOTSCH: *Matrix Browser: Visualisierung und Exploration vernetzter Informationsräume*, 2002

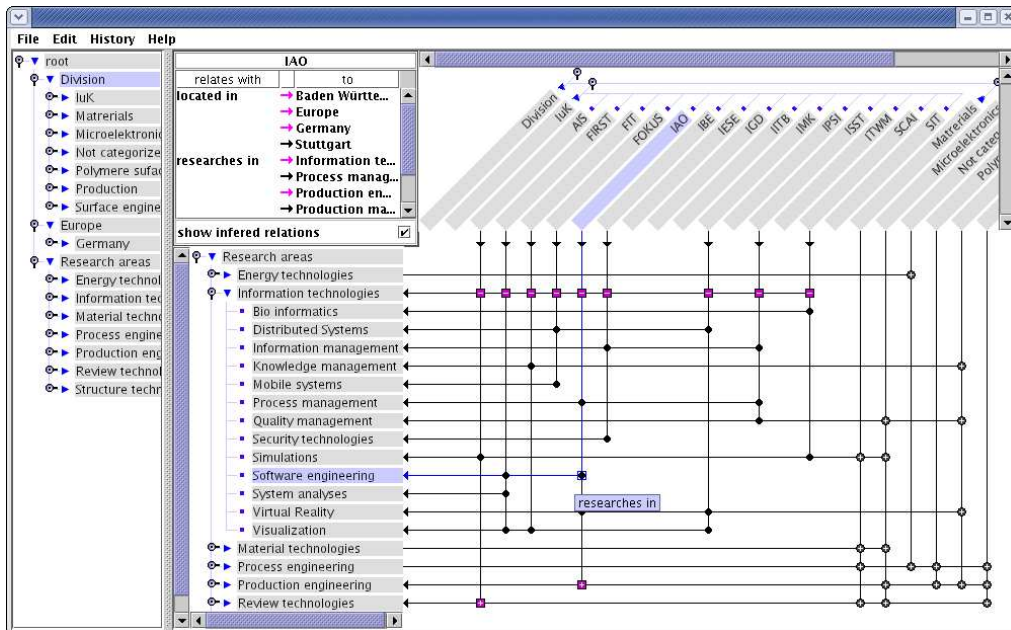


Abbildung 3.10: Prototyp des MatrixBrowsers

3.5.1.1 Grundidee: Matrixvisualisierung

Die zentrale Idee für die Konzeption des MatrixBrowsers ist es, den zugrunde liegenden Graphen auf eine hoch interaktive Darstellung einer Adjazenzmatrix abzubilden (siehe Abbildung 3.10). Die aus der Graphentheorie bekannten Adjazenzmatrizen sind eine alternative Form der Graphendarstellung. Die Knoten des Graphen werden an der horizontalen und vertikalen Achse der Matrix angeordnet, wobei die Zellen der Matrix die Kanten repräsentieren. Eine Zelle kennzeichnet eine Relation, wenn sich die Linien, die rechtwinklig zu den Achsen von zwei Knoten ausgehen, sich schneiden und diese beiden Knoten durch eine Kante verbunden sind. Durch Verwendung von Pfeilen und graphischen Symbolen können sowohl die Richtung der Relation als auch verschiedene syntaktische Typen von Relationen visualisiert werden. Unterstützend wird zusätzlich die Technik der ToolTips eingesetzt, um die unterschiedlichen textuellen Bezeichnungen der semantischen Beziehungstypen zu beschreiben.

Die zweite Grundidee ist es, diejenigen Teile des Ontologiegraphen, die eine hierarchische Struktur aufweisen, durch bekannte Baummetaphern darzustellen. Diese werden auf die beiden Achsen der Matrix platziert und können mittels vertrauten, "Windows Explorer"-ähnlichen Werkzeugen, effektiv erkundet werden, welche hierarchische Information visualisieren. Ontologiegraphen beinhalten typischerweise zahlreiche hierarchische Substrukturen, welche durch den semantischen Typ der einzelnen Relationen zustande kommen. Hierarchien werden zum Beispiel durch Subklassen-, Instanz- oder Aggregationsbeziehungen gebildet. Werden hierarchische Teilstrukturen bereits durch interaktive Bäume an den Matrixachsen dargestellt, wird die Markierung in den Zellen lediglich für weitere, zwischen den beiden Substrukturen bestehenden Relationen benötigt. Mit dem MatrixBrowser können die Knoten und Teilhierarchien, die an den Achsen dargestellt werden, auf flexible Art aus einer Obermenge ausgewählt und gefiltert werden, sowie durch schon gewohnte Expansion- und Kollabiermechanismen exploriert werden. So lassen sich zum Beispiel diejenigen Knoten auf einer beliebigen Achse herausfiltern, die nicht durch Relationen verbunden sind. Ebenfalls wird eine Sortierung der Knotenmengen anhand unterschiedlicher Kriterien wie die alphabetische Sortierung ermöglicht. Da eine Hierarchie auf den Achsen durch Subsumptions-, Aggregations- und Instanzbeziehungen der jeweiligen Knoten gebildet werden kann, ist ein weiteres Filterkriterium die Art dieser Beziehung oder eine Kombination derselben.

Diese Mechanismen ermöglichen es, die Information, welche in der Matrix dargestellt wird, besser zu strukturieren.

Durch die Verwendung dieser Baummetaphern sind nicht immer alle Knoten und Kanten sichtbar. Daher erlaubt der MatrixBrowser nicht nur das Expandieren und Kollabieren der Bäume, sondern auch der korrespondierenden Zellen. Wenn explizite Beziehungen gerade nicht sichtbar sind, weil die korrespondierenden Väterknoten sich in einem kollabierten Zustand befinden, wird ein interaktives Symbol in den Zellen angezeigt. Wird dieses selektiert, werden die damit assoziierten Knoten expandiert. Diese Techniken ermöglichen dem Nutzer sowohl ein systematisches Vordringen als auch eine Verdichtung der Informationsmenge.

Zusätzlich befindet sich eine tabellenbasierte, interaktive Visualisierung in der linken oberen Ecke der Matrix. Diese zeigt alle Relationen und direkten Nachbarn desjenigen Knoten im Netz, welcher in einer der beiden Bäume angewählt wurde. So ist nicht nur der Kontext des jeweiligen Knoten innerhalb einer taxonomischen Substruktur zu erkennen, sondern auch innerhalb des gesamten Netzes.

3.5.1.2 Interaktives Verhalten

Achsen

Die Untermenge der auf den Achsen dargestellten Knoten kann entweder durch die Teile des Netzes, welche hierarchische Eigenschaften besitzen, durch Setzen von Filtern oder durch Verkettung bestimmter Relationstypen (vgl. nachfolgenden Abschnitt 3.5.1.3) gebildet werden. Durch Expandieren und Kollabieren der Äste der jeweiligen Bäume können diejenigen Teile des Netzes, welche auf die Achsen gelegt wurden, direkt erkundet werden. Diese vertraute und effektive Art des Explorierens ist auf beiden Achsen möglich. Sind die jeweiligen Bäume zum Beispiel durch Expansion auf den Achsen zu groß, um innerhalb des Bildausschnitts dargestellt zu werden, erscheinen automatisch Schieberegler, mit denen das Darstellungsfeld verschoben werden kann.

Die beschriebenen Methoden, nutzerdefinierte Untermengen des Begriffsnetzes anzuzeigen, stellen flexible Möglichkeiten dar, den Informationsraum zu navigieren beziehungsweise visualisierte Ausschnitte davon zu vergrößern und zu verkleinern.

Matrix

Die in den Zellen der Matrix gezeigten Relationen zwischen den beiden, gerade auf den Achsen angezeigten Knotenmengen, sollten immer mit dem interaktiven Status der Achsen konsistent sein. Damit diese Konsistenz der Visualisierung gewährleistet ist, sind eine Reihe von Problemen zu lösen. Das erste Problem betrifft die Sichtbarkeit der Relationen von Knoten, die gerade in dem entsprechenden Baum nicht sichtbar sind, weil sich ein übergeordneter Ast in einem kollabierten Zustand befindet. Das zweite Problem tritt bei Relationen auf, bei denen die Beziehung zwischen zwei Knoten innerhalb der Hierarchie, die einen der beiden Knoten beinhaltet, nach oben oder unten weitergegeben werden kann.

Dieser Sachverhalt sei anhand eines Beispiels verdeutlicht: Innerhalb eines Beispiels, das die Organisationsstruktur einer Organisation mit den entsprechenden Standorten in Beziehung setzt, liegt der Standort einer Zweigstelle in einer Stadt. Dieses ist die explizite Verbindung im Netz. Der entsprechende Standort liegt aber auch in dem Land und auch innerhalb des Kontinents, in dem sich diese Stadt befindet. Daraus ergeben sich durch logisches Schließen zusätzliche Relationen. Ob und auf welcher Grundlage sich solche Inferenzen ergeben, hängt von den Axiomen und Regeln der Ontologie ab. Die Visualisierung jedoch ist auf diese Fälle ausgelegt.

Damit mit diesen unterschiedlich auftretenden Darstellungssituationen umgegangen werden kann, verfügt der MatrixBrowser über unterschiedliche Arten von visualisierten Relationstypen. Je nach Art der Beziehung verfügen auch die bildlichen Repräsentationen über interaktives Verhalten. Eine Relation in einer Zelle der Matrix kann dabei einen der folgenden Fälle repräsentieren:

- Eine *explizite* Relation stellt eine direkt spezifizierte Verbindung zwischen zwei Knoten des zugrunde liegenden Graphen dar. Diese Relation ist semantisch typisiert (z.B. verwandt-Mit).

- Eine *versteckte* Relation ist ein Indikator für eine weiter unten in der Hierarchie vorhandene explizite Relation, welche aber gerade nicht sichtbar ist, da sich der übergeordnete Ast in einem kollabierten Zustand befindet. Eine versteckte Relation trägt dabei keine weitere Bedeutung, sondern dient als Navigationshilfe.
- Eine *implizite* Relation zeigt eine Beziehung, welche durch den Deduktionsmechanismus aus der zugrunde liegenden Ontologie abgeleitet wurde. Diese kann entweder durch Vererbung oder Generalisierung entstehen. Bei Vererbung wird die Beziehung nach unten innerhalb der Hierarchie durchgereicht, während bei der Generalisierung die Weitergabe nach oben erfolgt. Speziell im zweiten Fall bedeutet diese Art von Beziehung, dass durch Expansion der entsprechenden Äste, die explizite Relation, welche der Grund für die Inferenz war, weiter unten in der Hierarchie gefunden werden kann.
- Eine *Identitätsrelation* stellt eine Beziehung zwischen zwei identischen Knoten her, welche aber in unterschiedlichen Hierarchien teilnehmen. Dieser Fall tritt auf Grund der multifacettierten Natur der Ontologie auf, da einzelne Knoten mehrere Väter besitzen können. Durch die Aufteilung der Ontologie in einzelne Hierarchien kann mit dieser Darstellungssituation umgegangen werden.

Die genannten Relationstypen werden visuell durch die Repräsentation mit unterschiedlichen Symbolen in den Zellen unterschieden. Im Falle der Generalisierung können sowohl versteckte als auch implizite Relationen, wie die Äste der Bäume, durch Mausklicks aufgeklappt werden. Damit diese Interaktionsmöglichkeit verdeutlicht wird, sind beide Arten durch ein "+" beziehungsweise "-" Symbol gekennzeichnet.

Eine enge Kopplung von Matrixzellen und Bäumen auf den Achsen erreicht, dass durch das Expandieren einer Relation auch eine Expansion der zugehörigen Äste erfolgt. Zusätzlich zu den beschriebenen Filterungs- und Sortiermechanismen der Knoten auf den Achsen, können Knoten auch auf Grundlage der Selektion von Relationen gefiltert werden. Dazu lässt sich ein rechteckiger Ausschnitt über die auszuwählenden Relation legen. Wird dieser Ausschnitt bestätigt, werden nur diejenigen Knoten (und Relationen) in der Matrix angezeigt, welche durch die ausgewählten Beziehungen verbunden sind.

Durch Selektion und Filterung des vollständigen Graphen und durch interaktives Herunterbrechen allgemeinerer Begrifflichkeiten auf speziellere Begriffe und ihren Relationen auf unterster Ebene, kann der Nutzer denjenigen Teil flexibel untersuchen, der gerade angezeigt wird.

Visualisierung von Knotennachbarschaften

Typischerweise werden einige visuelle Suchaufgaben von bestimmten Graphrepräsentationen bestens unterstützt, während andere nicht unbedingt von dieser Darstellung profitieren. Daher bietet der MatrixBrowser eine zusätzliche, alternative Visualisierung für denjenigen Fall, bei dem die Matrixdarstellung nicht optimal eingesetzt werden kann. Dieser Fall tritt auf, wenn die Ontologie nicht in genau zwei Teilbäume zerlegt werden kann. Typischerweise enthält eine Ontologie mehr als die gleichzeitig darstellbaren zwei Teilhierarchien.

Für diesen Fall wird eine tabellenbasierte Visualisierungstechnik zur Exploration der direkten Nachbarschaft von Knoten eingesetzt. Diese zeigt mittels dreier Spalten alle diejenigen Knoten an, welche mit dem auf einer Achse selektierten Knoten in Beziehung stehen. Die erste Spalte kennzeichnet dabei den semantischen Typ der jeweiligen Beziehung, während die zweite Spalte mit einem Pfeil die Richtung dieser Beziehung visualisiert. Dabei kann es sich um implizite oder explizite Relationen handeln. In der dritten Spalte steht einer der Knoten, welcher mit der vorangegangenen Beziehung und dem auf einer Achse angewählten Knoten verbunden ist. Dieser wird wiederum in der Überschrift der Tabelle angezeigt.

Wird in dieser Tabelle einer der Nachbarknoten selektiert, so ergibt sich daraus der neue Überschriftsknoten. Falls einer der Nachbarknoten nicht Teil der jeweils angezeigten Knotenmenge auf den Achsen der Matrix ist, wird er besonders markiert. Wird einer dieser speziellen Knoten in der Tabelle ausgewählt, werden gleichzeitig diese Knotenmengen oder Hierarchien in der Matrix angezeigt, in denen der selektierte Knoten Mitglied ist.

Ferner verfügt die Tabelle über eine Schaltfläche zur Filterung der angezeigten impliziten und expliziten Beziehungen. Die Schaltfläche kann an- oder abgewählt werden. Befindet sich diese in einem aktivierten Zustand, werden beide Arten der Beziehungen dargestellt. In einem passiven Zustand der Schaltfläche werden nur explizite Beziehungen in die Tabelle eingetragen.

Übersichtsfenster

Obwohl innerhalb der Matrix die vollständige Adjazenzmatrix dargestellt werden kann, das heißt alle Knoten sind sowohl auf der einen als auch auf der anderen Achse sichtbar, existiert eine Vielzahl an interaktiven Möglichkeiten, die dargestellte Informationsmenge zu reduzieren. Um aber ständig die Übersicht darüber zu behalten, was für Teilhierarchien und Knotenmengen in dem Netz enthalten sind, werden diese in dem Übersichtsfenster unterhalb eines abstrakten Wurzelknotens beziehungsweise unter dem allgemeinsten Begriff \top mittels der Baummetapher angezeigt. Von dort aus können sowohl einzelne Äste des Baumes als auch ganze Untermengen von Knoten via dem Mechanismus des Ziehen&Fallenlassens auf die einzelnen Achsen platziert werden.

Interaktionshistorie

Aufgrund der hohen Komplexität und Vielzahl der gebotenen Interaktionsmöglichkeiten bietet der MatrixBrowser eine Interaktionshistorie. Mit deren Hilfe lassen sich bereits getätigte Interaktionsschritte rückgängig machen und vorhergehende interaktive Zustände wieder herstellen. Als Erinnerungshilfe dieser Zustände wird bei jeder Veränderung des jeweiligen interaktiven Zustands ein bildlicher Schnappschuss des Matrix-Fensters gemacht und gespeichert. Durch diese Schnappschüsse können vorhergehende Zustände leichter erfasst werden.

3.5.1.3 Extraktion von Teilhierarchien

Damit die interaktiven Möglichkeiten des MatrixBrowsers optimal genutzt werden können, muss die zugrunde liegende Ontologie so aufbereitet werden, dass Teilstrukturen hierarchischer Natur extrahiert werden können. Nur hierdurch wird es möglich, semantisch zusammenhängende Teilstrukturen dynamisch zu filtern und zu visualisieren. Relationen, die eine Halbordnung über eine Knotenmenge legen, das heißt mit deren Hilfe eine Hierarchie dieser Knoten gebildet werden kann, sind hauptsächlich Subsumptions-, Aggregations- und Instanzbeziehungen. Je nach Aufgabenstellung ist es wünschenswert, diese drei Beziehungstypen verketteten zu können. Daher sind die jeweiligen Bäume auf den Achsen durch logische Anfragen beziehungsweise der Spezifikation von Rollenverkettung definiert. Eine solche Verkettung von Knoten lässt sich durch eine Verundung der Ausdrücke

$$\begin{aligned} C &\sqsubseteq C^* \\ i &: C \\ \exists(R_1 \circ R_2 \circ R_3 \circ \dots \circ R_n).C \end{aligned} \tag{3.8}$$

spezifizieren. Die Rollen $R_1 \dots R_n$ sind dabei neben den Subsumptions-, Aggregations- und Instanzbeziehungen beliebige Rollen der Ontologie, für die eine Transformation in Baumstrukturen gewünscht ist.

Werden im gesamten Graphen jeweils nur diese hierarchiebildenden Relationstypen ($\mathbf{R}_{>}$) betrachtet, ergeben sich Teilgraphen, die wiederum durch andersartige, in den Matrixzellen dargestellten Beziehungstypen ($\mathbf{R}_{matrix} | \mathbf{R} \notin \mathbf{R}_{matrix}$) verbunden sind. Dabei können sich die Teilgraphen überlappen, wenn sie gemeinsame Knoten besitzen. Diese Ausschnitte des Netzes werden durch eine vorhergehende Auflösung von eventuellen Zyklen in Baumstrukturen überführt. Innerhalb der zyklenfreien Baumstrukturen ist es immer noch möglich, dass ein Knoten mehrere Väter besitzt. Zur Darstellung können zwei Strategien verfolgt werden (vgl. Abbildung 3.11). Die eine Strategie dupliziert denjenigen Knoten, welcher multiple Väter besitzt und bildet daraus zwei eigenständige Hierarchien. Der verdoppelte Knoten ist dann in den Zellen der Matrix durch eine

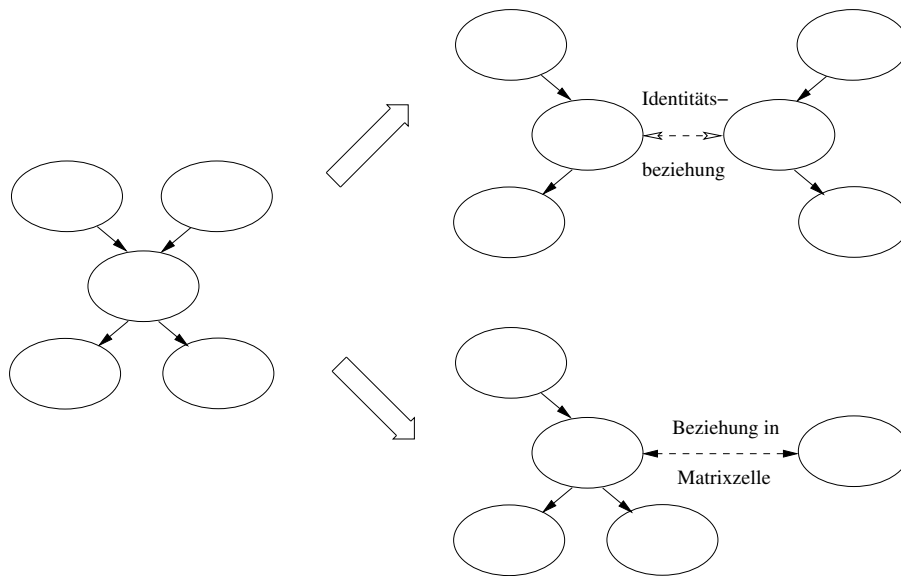


Abbildung 3.11: Strategien zur Extraktion von Teilhierarchien

Identitätsrelation gekennzeichnet. Die andere löst die Vaterschaftsbeziehung dadurch auf, dass nur einer der Väter zur Baumbildung berücksichtigt wird während die anderen Väter eigenständige Bäume bilden.

Mit der zweiten Variante der Baumbildung werden viele eigenständige Bäume erzeugt, die jeweils nur aus einem Knoten bestehen. Für eine übersichtliche Darstellung der Ontologie ist dies nicht wünschenswert. Daher ist für eine multifacettierte Ontologien die erste Variante vorzuziehen.

3.5.2 Visuelle Konstruktion semantischer Suchanfragen

Neben der Visualisierung einer Ontologie kommt der Hilfestellung bei der Konstruktion von logischen Abfragen eine besondere Bedeutung zu. Ohne eine solche Hilfe müssten die Abfragen in der Syntax der jeweiligen Deduktionskomponente erfolgen und könnten nicht von der Repräsentationssprache abstrahiert werden. Dies würde sowohl das Erlernen und eine genaue Kenntnis dieser Sprache als auch ein vorheriges Verständnis des Inhalts und der Bedeutung des ontologischen Modells erfordern. Außerdem müsste im Vorfeld bekannt sein, wie die einzelnen Klassen, Relationen und Attribute benannt wurden. In einem System mit der Anforderung der leichten Erlernbarkeit ist dies dem Nutzer kaum zumutbar. Existierende Methoden zur semantischen Abfrage wurden in Abschnitt 2.6.4, S. 64 diskutiert.

3.5.2.1 Einschränkung

Einer der Hauptgesichtspunkte bei der Gestaltung der visuellen Abfragesprache ist es, diese möglichst einfach zu halten, da auch unerfahrene Nutzer damit umgehen können sollten. Auf der anderen Seite sollten die damit formulierten Anfragen ausdrucksstark genug sein, um sämtliche Inhalte eines Informationsraums zu erreichen beziehungsweise abzufragen. So sollten alle syntaktischen Sprachkonstrukte von pOWL visuell ausgedrückt und die freien Variablen für interessante Teilausdrücke spezifiziert werden können (vgl. zur Syntax Abschnitt 3.3.2.1). Unter der vollständigen Berücksichtigung aller dieser Sprachkonstrukte würde die visuelle Abfragesprache jedoch eine zu hohe Komplexität erreichen, um auch für unerfahrene Nutzer anwendbar zu sein. Daher sind einige Abstriche auf der Seite der Ausdrucksfähigkeit der Abfragesprache zu machen. Folgende Einschränkungen werden gemacht:



Abbildung 3.12: Anfrageleiste zur Konstruktion semantischer Abfragen

- Es werden nur konjunktive Anfragen erlaubt ($C \sqcap D$).
- Zur Restriktion abstrakter Rollen werden nur Allquantoren erlaubt ($\forall R$ beziehungsweise $\forall R.C$ für $R \in \mathbf{R}_A$ und $C \in \mathbf{C}$).
- Zur Restriktion funktionaler Rollen (Attribute) werden ebenfalls nur Allquantoren erlaubt ($\forall T.C$ für $T \in \mathbf{R}_F$ und $C \in \mathbf{C}$).
- Es wird keine Angabe einer probabilistischen Randbedingung erlaubt.
- Anfragen werden nur für Individuen einer Ontologie evaluiert.

Ungeachtet dieser Einschränkungen lassen sich typische Suchkonstrukte durch die Verkettung dieser Ausdrücke formulieren. Ebenfalls können Attribute von Begriffsbeschreibungen oder Individuen (z.B. Nachname des Autors) restriktiert werden. Eine typische Suchanfrage wäre dazu "alle Dokumente des Typs X die von Thema Y behandelt werden, die wiederum mit Thema Z über Relation R verbunden sind".

3.5.2.2 Visuelle Abbildung

Das Hauptproblem bei der visuellen Konstruktion semantischer Abfragen besteht in der Auswahl von Begriffsbeschreibungen, Individuen und Rollen. Für diese Auswahl wird die Visualisierungstechnik des MatrixBrowsers herangezogen. Da in dieser Darstellung die relevanten Elemente der Ontologie auf den Achsen oder in den Matrixzellen dargestellt werden, stellt dies einen geeigneten Ausgangspunkt zur Auswahl dieser Elemente dar. Zur eigentlichen Konstruktion einer Anfrage ist eine spezielle Anfrageleiste mit der Matrixdarstellung so verbunden, dass sämtliche im MatrixBrowser visualisierten Ontologieelemente auf dieser Leiste über den Mechanismus des Ziehen&Fallenlassens platziert werden können. Ist auf eine solche Art die gewünschte Anfrage formuliert, kann die gesamte Anfragebeschreibung mit einem Druckknopf evaluiert werden. Ein weiterer Druckknopf löscht alle in der Anfrageleiste platzierten Objekte wieder heraus.

Die Anfrageleiste ist in Abbildung 3.12 dargestellt. Das Beispiel zeigt die visuelle Repräsentation der Anfrage "alle Informationsobjekte, die einen Autor besitzen". Diese Anfrage lässt sich in den logischen Ausdruck in pOWL-Syntax $x : (\text{InformationsObjekt} \sqcap \forall \text{creator.Autor})$ überführen, wobei x eine freie Variable und nach einer Evaluation diejenigen Individuen enthält, die Instanz der mit dem Ausdruck $\text{InformationsObjekt} \sqcap \forall \text{creator.Autor}$ beschriebenen Menge sind.

Die in der Anfrageleiste platzierten Ontologieelemente werden als Rechtecke dargestellt. Zusätzlich werden die jeweiligen lexikalischen Bezeichner der Ontologieelemente in einer ausgewählten Sprache (im abgebildeten Beispiel Englisch) angezeigt. Da Begriffsbeschreibungen durch funktionale Rollen attribuiert werden können, entsteht deren visuelle Repräsentation durch ein komplexeres Gebilde: Jede funktionale Rolle bildet durch ihren lexikalischen Bezeichner einen Eintrag in einer Auswahlliste. Innerhalb eines korrespondierenden Texteingabefeldes kann der tatsächliche Wert des Attributs, wie der Nachname eines Autors, eingegeben werden. Dabei hängt der einzugebende Wert von dem jeweiligen Datentyp ab, der mit der funktionalen Rolle verknüpft ist. Jede Eingabe eines Attributs entspricht einer weiteren Verkettung der Anfragebeschreibung mit $\sqcap \forall T.C = \text{Wert}$. Mit T ist dabei die funktionale Rolle und mit C die zugeordnete Begriffsbeschreibung bezeichnet.

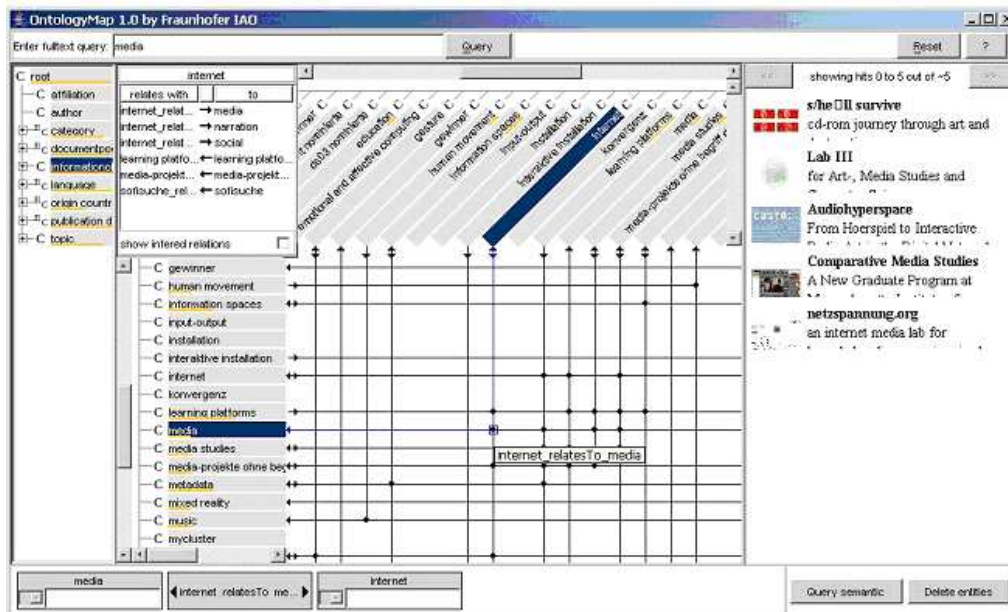


Abbildung 3.13: Schnittstelle zur Informationssuche

Zur weiteren Unterscheidung der in der Abfrageleiste platzierten Ontologieelemente werden Individuen nur durch ihren lexikalischen Bezeichner innerhalb des Rechtecks ausgewiesen, während Rollen durch einen zusätzlichen Pfeil visualisiert werden. Die Richtung des Pfeils zeigt dabei auf dasjenige Individuum oder Begriffsbeschreibung, welches zum Definitionsbereich der jeweiligen Rolle gehört. Jede nach einer Begriffsbeschreibung oder einem Individuum angeordnete Rolle entspricht einer weiteren Verkettung der Anfragebeschreibung mit $\forall R.C$, wobei sich C auf das nachstehende Ontologieelement (Begriffsbeschreibung oder Individuum) bezieht.

Die Anordnung der Ontologieelemente ist dabei beliebig. Jede aufeinander folgende Begriffsbeschreibung entspricht der Verkettung der Anfrage mit $\exists C \sqcap D$. Jede Kombination aus einer Begriffsbeschreibung mit einer Rolle entspricht der weiteren Konjunktion der Anfragebeschreibung mit $\forall R.C$. Jede doppelte Anordnung einer Rolle entspricht der Konjunktion mit $\forall R.(\forall R.C)$. Es sei darauf hingewiesen, dass die sinnvolle Konstruktion einer Anfrage aufgrund dieser Freiheitsgrade nicht sicher gestellt werden kann. Zum Beispiel lässt sich die Anfrage $C \sqcap \{i\}$ formulieren. Dies entspricht einer Konjunktion einer Begriffsbeschreibung mit einem Individuum, was in jedem Fall zu einer leeren Ergebnismenge führt.

3.5.3 Integration der Techniken

Die beschriebenen Techniken zur bildlichen Präsentation ontologischer Navigation und zur visuellen Konstruktion semantischer Anfragen sind mit einer zusätzlichen Abfragemöglichkeit in einer einheitlichen Nutzungsschnittstelle integriert. Abbildung 3.13 zeigt die prototypische Umsetzung dieser Ideen. Das Hauptfenster der Anwendung ist dabei in eine herkömmliche Arbeitsplatsumgebung eingebettet, wie sie moderne Betriebssysteme mit graphischen Oberflächen bieten. Es teilt sich dabei in drei wesentliche Bereiche auf: Am oberen Fensterrand befindet sich ein Texteingabefeld mit einem Anfrageknopf zur Formulierung und Durchführung von stichwortbasierten Suchanfragen. Der mittlere Bereich ist für die Visualisierung der Ontologie auf Basis des Matrixbrowsers (links) und für die Präsentation der Suchergebnisse (rechts) vorgesehen, während am unteren Rand des Anwendungsfensters ein Feld zur visuellen Konstruktion und Ausführung semantischer Anfragen platziert ist.

3.5.3.1 Aufbereitung der Ontologie

Die gesamte recherchierbare Ontologie besteht im Minimalfall aus einer semantischen Beschreibung von Dokumenten auf Grundlage der in Abschnitt 3.6 beschriebenen Minimalabstraktionen. Zusätzlich lassen sich, wie gezeigt, in dieser Struktur weitere Arten von Inhalten wie zum Beispiel Personaldaten, Organigramme oder Produktdaten einbetten, wenn diese durch eigene Ontologien modelliert wurden. Aus diesem Grund kann die systemische Wissensbasis sehr große Dimensionen annehmen. Eine umfassende Darstellung aller Individuen und Begriffsbeschreibungen mit ihren Beziehungen ist daher nicht wünschenswert.

Daher werden nur Begriffsbeschreibungen mit deren Subsumptionsbeziehungen und Rollen zur Visualisierung herangezogen. Eine Ausnahme dieser Regelung bilden Begriffsbeschreibungen, die ohne ihre Individuen keine vollständige Ausdruckskraft besitzen. Ein Beispiel hierzu wäre die Begriffsbeschreibung des Herkunftslandes eines Dokuments. Die einzelnen Herkunftsländer müssen zwangsläufig als Individuen modelliert sein. Da es sich dabei um eine überschaubare Menge an Individuen handelt, werden diese in die Ontologievisualisierung mit einbezogen und erleichtern damit auch eine entsprechende Abfrage. Solche Begriffsbeschreibungen lassen sich anhand ihrer speziellen Definitionsweise, der Aufzählung von Individuen mittels $C \doteq \{i_1, i_2, \dots, i_n\}$ identifizieren.

3.5.3.2 Kopplung von Matrix- und Tabellendarstellung

Anhand dieser Aufbereitungsstrategie lässt sich eine Trennung von Klassen- und Instanzebene vornehmen. Während im MatrixBrowser die schematische Ebene dargestellt ist, werden in einer tabellenbasierten Präsentationsform nur Individuen der Ontologie visualisiert. Die tabellenbasierte Form der Präsentation beruht auf einer Studie von Sebrechts et al.²⁸. In dieser Studie wurde der tatsächliche Wertvorteil von unterschiedlichen Suchergebnispräsentationen des Information Retrievals untersucht. Verglichen wurden dabei Visualisierungstechniken in allen drei Dimensionen. In der Untersuchung stellte sich heraus, dass die allgemein gebräuchliche flache Texttabelle den anderen Techniken vorzuziehen ist, wenn Vorwissens und Erfahrung der Nutzer berücksichtigt werden.

Die auf Schemaebene der Ontologie operierende Matrixvisualisierung und die Tabelle, welche die zugehörige Individuenmenge präsentiert, sind nach dem Prinzip der direkten Manipulation gekoppelt. Mit Hilfe einer speziellen Art der Knotenselektion, dem Doppelklick mit der Maus, werden die Individuen der korrespondierenden Begriffsbeschreibungen abgefragt und in der Tabelle angezeigt. Auch hier existiert eine Ausnahme: Handelt es sich bei der selektierten Begriffsbeschreibung um Unterbegriffe des Gegenstandsbereich der Minimalabstraktionen, dann sind dort nicht nur Dokumente, sondern auch eventuelle andere Inhalte klassifiziert (vgl. Abbildung 3.7, S. 89). Da aber Dokumente im hauptsächlichen Interesse eines IR-Systems stehen, erfolgt in diesem Falle zur Abfrage der Individuen die konjunktive Begriffsbeschreibung $x : C \sqcap \text{Dokument}$ im Gegensatz zur standardmäßigen Abfrage $x : C$. Die freie Variable x wird hierbei an die Individuen der Begriffsbeschreibung gebunden, während C die korrespondierende Begriffsbeschreibung zu dem jeweiligen doppelgeklicktem Knoten repräsentiert.

Der tatsächliche Inhalt der mit der Tabelle dargestellten Individuen oder Suchergebnissen, lässt sich durch die Selektion eines Tabelleneintrags per Doppelklick mit der Maus einsehen. Es wird dann ein weiteres Fenster geöffnet, welches eine detaillierte Beschreibung des jeweiligen Suchergebnisses zeigt.

3.5.3.3 Kontextualisierung von Suchergebnissen

Die Kombination aus Strukturvisualisierung und Individuenmenge lässt sich für eine thematische Kontextualisierung von Ergebnissen einer Suchanfrage heranziehen. Dies ist besonders bei stichwortbasierten Suchanfragen vorteilhaft, da dort von der Stichwortsuchmaschine aufgrund

²⁸SEBRECHTS et al.: *Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces*, 1999

von gleichen Stichwörtern, Dokumente unterschiedlicher Themenzugehörigkeit zurückgeliefert werden können (vgl. Abschnitte 3.13, S. 106).

Wird eine Suchanfrage evaluiert, werden die Ergebnisse der Suche in der Tabelle dargestellt. Für jedes einzelne Suchergebnis wird gleichzeitig, durch Anfrage der Wissensbasis die Menge der Klassen beziehungsweise Begriffsbeschreibungen nach $\{i_i\} : x$ ermittelt, in welchen das Suchergebnis ein Individuum darstellt. Für jede Begriffsbeschreibung wird ein Zähler inkrementiert. Nach der Bearbeitung aller Suchergebnisse enthält so der allgemeinste Begriff \top die Anzahl aller Suchergebnisse. Die Knoten, welche den Begriffsbeschreibungen entsprechen, werden dann in der Matrix mit einem Balken markiert. Die Länge des Balkens gibt das Verhältnis von den, in dieser Begriffsbeschreibung gefundenen Treffern, zu der Gesamtzahl aller Treffer an. Auf diese Weise erhält der Nutzer eine Rückkopplung, wie die erzielten Treffer einer Suche mit der Anfrage korrelieren.

In diesem interaktiven Zustand der Anwendung lässt sich dann die Menge aller Suchergebnisse über den Bezug zu Begriffsbeschreibungen der Ontologie einschränken. Wird ein Knoten in der Matrix selektiert, so werden nur diejenigen Treffer in der Tabelle angezeigt, die Mitglied der Menge sind, die mit dem angewählten Knoten beschrieben werden.

3.5.3.4 Dreistufiger Rechercheprozess

Neben der explorativen Erschließung der Wissensbasis mittels des MatrixBrowsers und der damit verbundenen Tabelle wird unter Verwendung der Gesamtheit der Interaktions- und Visualisierungstechniken, ein Rechercheprozess in drei Schritten angeboten. Innerhalb dieser drei Schritte kann ein bestimmter Informationsbedarf iterativ präzisiert werden.

1. Um die für ein bestimmtes Informationsbedürfnis relevanten Ausschnitte des visualisierten Netzes zu finden, beginnt der Nutzer mit der Eingabe von Stichwörtern. Hierauf vollzieht das System eine standardmäßige Stichwortsuche, wobei nicht nur die reine Ergebnismenge als Tabelle dargestellt wird (vgl. Abbildung 3.13, rechts), sondern auch die gefundenen Dokumente den Begriffen in der Ontologie zugeordnet werden. Diese Begriffe werden markiert und kontextualisieren die Treffermenge. Sämtliche auf diese Art gefundenen Begriffshierarchien können beliebig via Ziehen&Fallenlassen auf die Matrixachsen gezogen und damit exploriert werden.
2. Im zweiten Schritt kann die Ergebnismenge dadurch weiter eingeschränkt werden, dass einzelne Begriffe auf den Matrixachsen selektiert werden. Es ergibt sich die Schnittmenge aus allen Trefferdokumenten und denen, die unter einer Begriffsbeschreibung als Individuen aufgehängt sind. Auch können alle direkten Individuen, das heißt Dokumente, eines Begriffs per Doppelklick auf einen Themenknoten angezeigt werden.
3. Im dritten Schritt können die in der Ontologie enthaltenen Metadaten (Autor, Erscheinungsjahr etc.) und thematische Oberbegriffe, sowie die expliziten und impliziten Relationen dazu verwendet werden, tatsächliche semantische Abfragen anhand der logischen Struktur zu generieren. Hierzu existiert die spezielle Anfrageleiste, auf die wiederum mittels Ziehen&Fallenlassen gewünschte Ontologieelemente und Relationen gezogen werden können. So wird es möglich die Ontologie im Sinne von "alle Dokumente des Typs X die von Thema Y behandelt werden, die wiederum mit Thema Z über Relation R verbunden sind" abzufragen.

4 Evaluierung

Im diesem Kapitel werden die Ergebnisse einer Evaluation des oben vorgestellten Ansatzes präsentiert. Die dazu durchgeführten Untersuchungen bezogen sich auf das Verfahren der kollaborativen Wissensakquise, sowie auf die Entwicklung der Interaktions- und Visualisierungswerkzeuge und deren Integration in die Recherveschnittstelle. Entsprechend den formulierten Zielen dieser Arbeit richten sich die Bewertungen der einzelnen Auswertungen an den Bedürfnissen typischer Nutzer.

Auf der Verfahrensseite der kollaborativen Wissensakquise wurde eine Studie zu Ablagestrukturen von zwei unterschiedlichen Personengruppen zu unterschiedlichen Dokumentbeständen durchgeführt. Diese hatte zum Ziel die grundlegenden Hypothesen und damit die prinzipielle Anwendbarkeit eines solchen Verfahrens zu überprüfen, sowie die Qualität der generierten probabilistischen Ontologien im Hinblick auf die Benutzbarkeit im vorgestellten Rahmenwerk festzustellen.

Die Bewertungen auf der gestalterischen Ebene der Mensch-Computer Interaktion teilten sich in zwei wesentliche Bereiche auf: So erfolgte zum einen die Entwicklung der Strukturvisualisierung mit dem MatrixBrowser als zentralem Element der Recherveschnittstelle innerhalb eines iterativen und experimentell gestützten Vorgehens nach ISO 13407¹. Diese internationale Norm definiert einen benutzerorientierten Gestaltungsprozess bei der Entwicklung rechnergestützter interaktiver Systeme zur Sicherstellung deren Gebrauchstauglichkeit. Jede Iteration dieses Prozesses bedingt Evaluationsstudien als Ausgangspunkt für den nächsten Durchlauf. Zum anderen wurde die Recherveschnittstelle in ihrer Gesamtheit einer umfangreichen Untersuchung unterzogen. Dabei stand im Vordergrund, inwieweit die an den Rechercheansatz gestellten Anforderungen, wie die Effektivität der Zugangs- und Abfragemechanismen, der Förderung des Verständnis von Informationsraum und Informationsbedarf, sowie der inhaltlichen Erfassung der Ergebnismenge einer Suchanfrage erfüllt sind.

4.1 Evaluierung des Verfahrens zur kollaborativen Wissensakquise

Das Verfahren des kollaborativen Indexierens dient zur möglichst automatisierten initialen Erstellung und Pflege von probabilistischen Themenontologien. Diese werden durch statistische Analyse und Überlagerung von Informationsablagestrukturen wie Lesezeichen oder Dateisysteme gebildet, welche entweder Verweise auf genutzte Inhalte oder die Inhalte selbst umfassen. Zur Bewertung der Qualität und der Anwendbarkeit dieses Verfahrens innerhalb des hier verfolgten Ansatzes wurde eine Evaluationsstudie durchgeführt. Innerhalb der Studie wurden Ontologien aufgrund von Ablagestrukturen untersucht, die in zwei Szenarien entstanden. Im ersten Szenarium wurden die Ablagestrukturen einer heterogenen Personengruppe für einen offenen Informationsraum (Internet) erzeugt, während sie im zweiten Szenario von einer homogenen Personengruppe für einen abgeschlossenen Informationsraum (Projektdatenbank) gebildet wurden.

4.1.1 Ziel und Fragestellung

Die erfolgreiche Pflege beziehungsweise Erstellung von neuem Systemwissen aus Informationsablagestrukturen innerhalb des Verfahrens zum kollaborativen Indexieren stützt sich auf die in Abschnitt 3.4.1, S. 92 formulierten Hypothesen und Annahmen. Das Hauptziel der Studie war die Verifikation dieser Hypothesen, das heißt inwieweit Informationsablagestrukturen den gemachten Annahmen entsprechen, so dass valides Wissen aus diesen extrahiert werden kann.

¹INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: *ISO 13407: Human-centred Design Processes for Interactive Systems*, 1999

Ein weiteres Ziel war die Qualitätsbeurteilung der durch das Verfahren generierten Ontologien zur Bewertung von dessen Anwendbarkeit. Folgende Fragestellungen standen bei der Auswertung im Vordergrund:

Qualität der atomaren Begriffsbeschreibungen: Wie hoch ist die allgemeine Verständlichkeit der aus den Ordnungsbezeichnern entstandenen atomaren Begriffsbeschreibungen? Insbesondere sind dabei solche Begriffe wertvoll, die spezifisch genug Themen beschreiben oder für eine sonstige Bedeutsamkeit einer Menge von Inhalten stehen. Diesen gegenüber stehen zum Beispiel umgangssprachliche, orthographisch falsche Wörter, Begriffe aus der Mundart und Abkürzungen, welche im Sprachgebrauch unüblich sind. Für die Nutzung der probabilistischen Ontologie zur Navigation, Exploration und Suche des zugrundeliegenden Informationsraums sind diese letzteren, im Hinblick auf verständliche Begrifflichkeiten für den Nutzer, ungeeignet.

Qualität der Begriffsassertionen: Begriffsassertionen werden aus den, in benannten Ordnern abgelegten Lesezeichen oder Dateien gebildet. Wie hoch ist dabei die Güte der inhaltlichen Zuordnung einer Menge von Individuen zu ihrer Begriffsbeschreibung beziehungsweise des inhaltlichen Zusammenhangs dieser Menge?

Qualität der Rollen: Beziehungen zwischen Begriffsbeschreibungen werden durch gemeinsame Informationsobjekte in verschiedenen Ordnern hergestellt. Sind die so entstandenen Rollen von Begriffsbeschreibungen inhaltlich gültig?

Qualität der Begriffsinklusionen: Subsumtionsbeziehungen von Ober- und Unterordnern innerhalb von Ablagestrukturen werden durch eine statistische Auswertung in Begriffsinklusionen überführt. Sind diese hierarchischen Beziehungen gültig?

4.1.2 Vorgehensweise

Als heterogene Personengruppe in dem ersten Szenarium wurden Mitarbeiter des Fraunhofer Instituts für Arbeitswirtschaft und Organisation² (IAO) herangezogen. Die Ausgangsdaten für das kollaborative Indexieren stellten dabei Lesezeichen von Internetbrowsern des täglichen Gebrauchs dar. Die Lesezeichen unterlagen keinen Beschränkungen. Sie konnten entweder flach oder auch in (hierarchischen) Ordnern organisiert sein. Die Lesezeichen verwiesen auf Webseiten des gesamten Internets, welches den offenen Informationsraum repräsentiert. Über einen elektronischen Rundbrief an alle Mitarbeiter des IAO (inklusive der wissenschaftlichen Hilfskräfte) wurden diese gebeten, ihre Browserlesezeichen für die Studie zur Verfügung zu stellen. Dazu wurde eine Webseite zur Verfügung gestellt, welche das einfache und anonymisierte Hochladen der Lesezeichen erlaubte. Nutzer des InternetExplorers³ konnten dies über einen Schalter auf der Webseite initiieren, welcher mit einem speziellen Skript hinterlegt war, das den Export und das Hochladen auf die Webseite automatisierte. Als Datenformat wurde das Netscape-Format von Lesezeichen gewählt, welches von allen Browsertypen unterstützt wird. Wurden andere Browsertypen als der InternetExplorer verwendet, so mussten die Lesezeichen im Netscape-Format in eine Datei exportiert und diese hochgeladen werden. Serverseitig wurden die Lesezeichen ebenfalls im Netscape-Format in einer Datei gespeichert und standen so zur weiteren Verarbeitung zur Verfügung.

Das zweite Szenarium entstand innerhalb des, vom Bundesministerium für Bildung und Forschung⁴ (BMBF) geförderten, Forschungsprojekts AWAKE⁵ - Networked Awareness for Knowledge Discovery - am MARS Exploratory Media Labs. Das MARS Exploratory Media Lab war

²<http://www.iao.fraunhofer.de>

³Internetbrowser der Firma Microsoft GmbH; <http://www.microsoft.com/germany/ms/internet/default.htm>

⁴<http://www.bmbf.de>

⁵NOVAK/KUNZ/WURST: *Entdeckung und Nutzbarmachung von stillem Wissen in heterogenen Expertengemeinschaften*, 2003

einer der Projektpartner innerhalb des AWAKE-Projekts und ist eine Fachabteilung des Fraunhofer Instituts für Medienkommunikation⁶ (IMK) dar. In dem Projekt AWAKE wurde die Fragestellung untersucht, wie und inwieweit bereits bestehende, aber noch nicht explizit formulierte Wissensstrukturen einer bestimmten Expertengemeinschaft entdeckt, visualisiert und für die kooperative Entdeckung und Konstruktion von neuem Wissen in heterogenen Informationsräumen nutzbar gemacht werden können. Als konkreter Untersuchungsrahmen wurde der Kontext der Suche nach Information und der Exploration eines Informationsraums gewählt. Dieser Kontext kann als ein Prozess verstanden werden, in dem sich vorhandenes Wissen der Suchenden durch Interaktion mit einzelnen Inhalten des Informationsraums widerspiegelt und dadurch neue Wissensstrukturen entwickelt werden. Um die Wissensstrukturen des einzelnen Nutzers in Bezug zu einem Informationsraum zu stellen und um einen Kontext zur Externalisierung derselben zu bieten, wurde die Metapher einer Wissenskarte heran gezogen. Diese Wissenskarten wurden als eine interaktive Visualisierung der thematischen Struktur eines Informationsraums innerhalb eines Computersystems implementiert. Dabei sind thematisch zusammengehörige Dokumente in benannten Gruppen zusammengefasst, welche dann als zweidimensionales Gebiet auf einer Wissenskarte dargestellt werden. Die einzelnen Dokumente sind als Punkte in diesen Gebieten repräsentiert. Ausgehend von der initialen Strukturierung des Informationsraums, welche autonom durch das System mittels eines neuronalen Netzwerks hergestellt wird, erhalten die Nutzer die Möglichkeit sowohl den Informationsraum explorativ zu erkunden, als auch persönliche Wissenskarten anzulegen. Die Erstellung persönlicher Wissenskarten geschieht durch das Umordnen der systemgenerierten Struktur, in dem zum Beispiel relevante Informationsobjekte ausgewählt und Objektgruppen zugeordnet, Objekte zwischen Objektgruppen bewegt oder ganze Gruppen neu erstellt werden. Der Vorgang des Erstellens einer solcherartigen persönlichen Wissenskarte kann mit dem Anlegen eines zweidimensionalen Lesezeichens der thematischen Gruppen verglichen werden. Allerdings besteht hierbei die Einschränkung, dass diese Lesezeichen zwar in Gruppen, nicht aber in hierarchischen Ordnerstrukturen angelegt werden können.

Prototypisch wurde das AWAKE System (unter anderem) für den Informationsraum von "netzspannung.org"⁷ implementiert. "netzspannung.org" strebt an, ein Wissensportal einzurichten, das Einblick in die Schnittpunkte von digitaler Kunst, Kultur und Informationstechnologie bietet. Es umfasst wissenschaftliche Veröffentlichungen und Projektbeschreibungen sowohl im kommerziellen als auch im Forschungsumfeld. Ebenfalls werden mit der "Digital Sparks" genannten Ausschreibung innerhalb des Portals studentische Projekte gesucht und die Gewinner dieser Ausschreibung gefördert.

In diesem zweiten Szenarium repräsentierten die zwölf Mitglieder des MARS Exploratory Media Labs die homogene Nutzergruppe. Diese wurden aufgefordert, Wissenskarten für den abgeschlossenen Informationsraum von "netzspannung.org" zu erstellen. Ziel dabei war es, persönlich relevante Projektbeschreibungen thematisch zusammenzufassen, so dass die entstehenden Wissenskarten neuen Nutzern von "netzspannung.org" im Sinne eines geführten Rundgangs vorgelegt werden könnten.

Die in den beiden Testszenarien entstandenen Ablagestrukturen wurden mittels dem Verfahren des kollaborativen Indexierens in zwei probabilistische Ontologien überführt. Für jede dieser Ontologien wurde eine bereinigte Ontologie erzeugt, welche nur Begriffsbeschreibungen enthielt, die mindestens zweifach als Ordnungsbezeichner in Ablagestrukturen vorkamen. Zur Überprüfung der statistischen Filterbarkeit von unscharfen sowie ungewöhnlichen Begriffen wurden diese Ontologiepaare getrennt betrachtet.

Da innerhalb der Nutzerschnittstelle des AWAKE-Systems keine Subsumtionsbeziehungen der Lesezeichenordner angelegt werden konnten, entfiel die Qualitätsbeurteilung der Begriffsinclusionen im Rahmen des zweiten Szenarios.

⁶<http://www.imk.fraunhofer.de>

⁷FLEISCHMANN/STRAUSS/ET AL: *netzspannung.org: an Internet Media Lab for Knowledge Discovery in Mixed Realities*, 2001; siehe auch: <http://www.netzspannung.org>

4.1.3 Ergebnisse

4.1.3.1 Browserlesezeichen für einen offenen Informationsraum

In der ersten Studie stellten 24 Mitarbeiter des Fraunhofer Instituts für Arbeitswirtschaft und Organisation ihre Browserlesezeichen der Auswertung zur Verfügung. Durch die Auswertung der 24 Ablagestrukturen entstanden insgesamt 515 Begriffsbeschreibungen. Darin waren sechs Begriffsbeschreibungen wie "*medium*" oder "*Ordner persönliche Symbolleiste*" enthalten, welche durch voreingestellte Ordner innerhalb spezifischer Internetbrowser entstanden. Diese wurden in der weiteren Auswertung nicht betrachtet. Die ausgewerteten Ablagestrukturen enthielten 4708 Lesezeichen, von denen 3720 in Ordnern organisiert und 988 lose abgelegt waren. Das entspricht einem Verhältnis von 79% geordneten zu 21% ungeordneten Lesezeichen. Durch Strukturvergleich der Ordner wurden 162 Rollen für die, zu den Ordnerbezeichnern korrespondierenden, Begriffsbeschreibungen generiert. Von den verbleibenden 509 Begriffsbeschreibungen sind 273 (54%) in Hierarchien angeordnet, während 236 (46%) auf oberster Hierarchieebene stehen. Dabei entstanden Hierarchietiefen von bis zu sechs Ebenen.

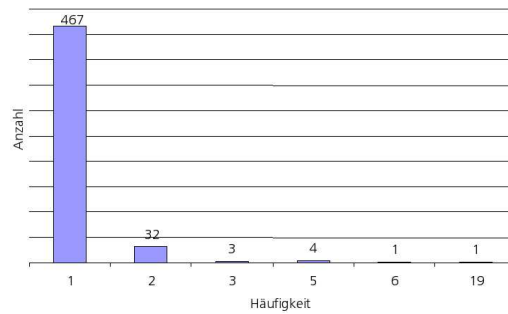
Durch das Filtern von Begriffsbeschreibungen mit einer Häufigkeit von eins mit ihren zugehörigen Rollen und Individuen wurde eine bereinigte Ontologie mit 42 verbleibenden Begriffsbeschreibungen erzeugt. Durch diese werden 642 Individuen beschrieben. Aufgrund des Filtervorgangs verblieben drei Rollen in der bereinigten Ontologie. Von den 42 Begriffsbeschreibungen sind 16 in Hierarchien und 26 auf oberster Ebene angeordnet. Die maximale Hierarchietiefe beträgt drei Ebenen.

Die absoluten Häufigkeiten der entstandenen Begriffsbeschreibungen zeigt Abbildung 4.1 (a). Die häufigsten Ordnungsbezeichner sind "*links*"⁸ mit 19fachem Vorkommen, "*privat*" mit sechsfachem Vorkommen, sowie "*Projekte*", "*3D*", "*Java*" und "*XML*" mit jeweils vierfachem Vorkommen. Die verwendeten Begrifflichkeiten zerfallen in 396 Wörter für Themengebiete wie "*Mathematik*" oder "*Kunst*" und in 113 sonstige Wörter, welche entweder ungebräuchlich oder zu unspezifisch für eine klar umrissene Klassifikation sind. Dies entspricht einem Anteil von 78% der Themenbegriffe beziehungsweise 22% der sonstigen Wörter innerhalb der ontologischen Begriffsbeschreibungen (vgl. Abbildung 4.1 (b)). Die Gesamtheit der Themenbegriffe wurde aus 347 Wörtern (68%) der deutschen Sprache und aus 49 (10%) Wörtern der englischen Sprache gebildet. Die Menge der sonstigen Begriffe konnte in 93 ungebräuchliche Wörter, welche wie "*Muziq*" oder "*Filme guggen*" der Umgangssprache und Mundart entlehnt sind, sowie in 20 unscharfe Begriffe wie "*Diverses*" oder "*Privat*" unterteilt werden. Dies entspricht einem Anteil von 18% ungebräuchlicher und vier Prozent unscharfer Begriffe. Abbildung 4.1 (c) zeigt die absoluten und prozentualen Anteile der obigen Wortkategorien in der bereinigten Ontologie. Der Anteil der 35 Themenbegriffe beträgt nun 83%, welche von 30 Wörtern (71%) in deutscher Sprache und von 5 Wörtern (12%) in englischer Sprache gebildet werden. Der Bereich der sieben (17%) sonstigen Wörter besteht aus einem ungebräuchlichen Wort (zwei Prozent) und 6 unscharfen Wörtern (15%).

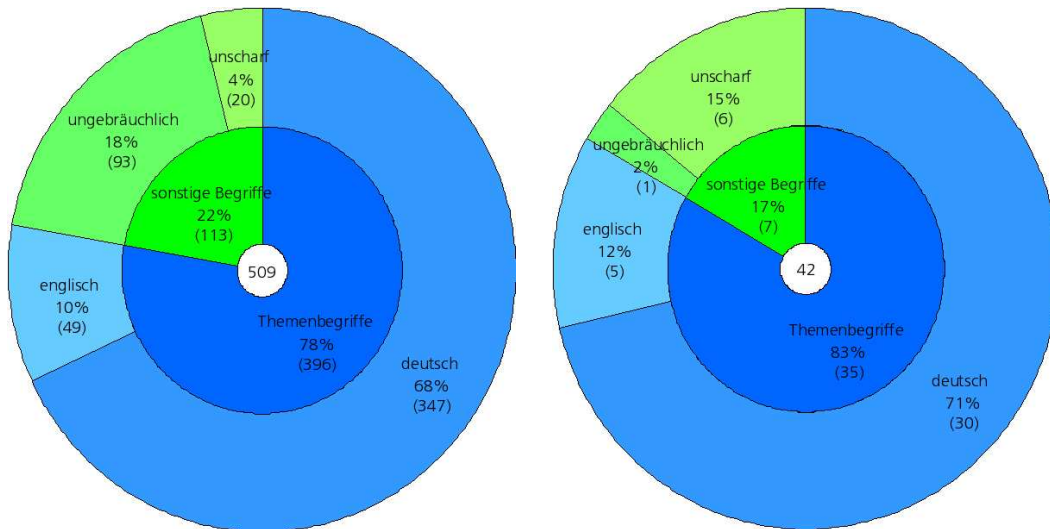
Von den 3720 entstandenen Individuen sind 161 multifacettiert und 3559 monofacettiert von Begriffsbeschreibungen klassifiziert. Dies bedeutet einen Anteil von vier Prozent der Individuen mit mehr als einer zugehörigen Begriffsbeschreibung. Dem gegenüber stehen 96% von Individuen mit genau einer klassifizierenden Begriffsbeschreibung. Eines der Individuen mit den meisten Facetten ist die Webseite "<http://www.teleauskunft.de>", der Übersichtsseite zu dem Telefonbuch, Gelbe Seiten und dem örtlichen Telefonbuch des Auskunftssystems der Deutschen Telekom Medien GmbH. Diese wird durch die Begriffe "*Post Telefon*", "*Telefon und Adresse*", "*Datenbanken*" und "*Benutzen*" beschrieben. In der 642elementigen Individualmenge der bereinigten Ontologie befinden sich drei (0,5%) multifacettierte und 638 (99,5%) monofacettierte Individuen. Hierin ist eines der Individuen mit den meisten Facetten die Webseite "<http://selfhtml.teamone.de>", einer Lehrseite zu HTML und JavaScript. Diese wird durch die Begriffe "*Javascript*" und "*Html*" beschrieben.

Aufgrund der großen Menge von Klassen wurde die Qualitätsbeurteilung der inhaltlichen Zuordnung von Individuen zu Begriffsbeschreibungen auf die Menge der 35 Themenklassen innerhalb der bereinigten Ontologie begrenzt. Von den damit beschriebenen 539 Individuen sind 533

⁸im Sinne eines Verweises, keine Richtungsangabe



(a) Häufigkeitsverteilung der Begriffsbeschreibungen



(b) Wortkategorien (unbereinigte Ontologie)

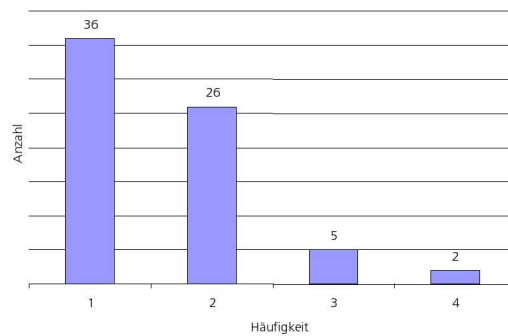
(c) Wortkategorien (bereinigte Ontologie)

Abbildung 4.1: Statistiken von Begriffsbeschreibungen nach der Auswertung von Browserlesezeichen

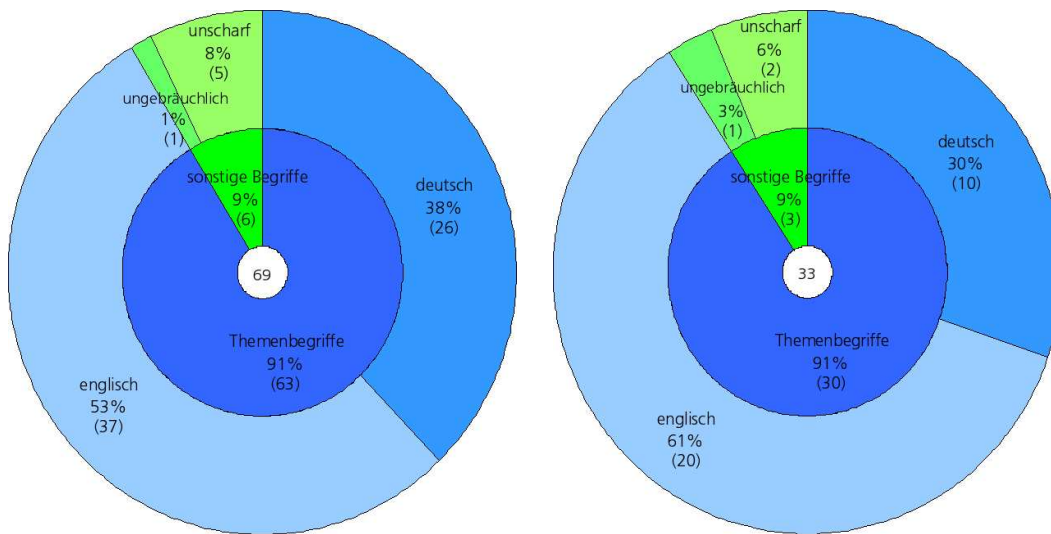
korrekt und 6 falsch klassifiziert. Dies entspricht einer korrekten Klassifikationsquote von 98,8%.

Von den 162 generierten Rollen konnten 161 als semantisch korrekt und eine als falsch eingestuft werden. Dies entspricht 99,3% semantisch korrekter Rollen. So steht zum Beispiel der Begriff "Wissen" zu "Bildung" und "Wissenschaft" in einer korrekten semantischen Verbindung. Die einzige semantisch falsche Relation ist diejenige, welche zwischen "Job" und Geschichte der "frühen Neuzeit" gebildet wurde. Werden nur Rollen aus der bereinigten Ontologie betrachtet, so sind diese semantisch einwandfrei.

Für die Qualitätsbeurteilung von Subsumtionsbeziehungen wurden die Pfade von Begriffen der untersten Hierarchieebene bis zu ihrem Wurzelknoten der obersten Ebene herangezogen. Eine Betrachtung der so entstandenen 302 Subsumtionspfaden ergab 206 (98%) korrekte und sechs (zwei Prozent) falsche Pfade. Der längste korrekte Pfad war hierbei "thematisch geordnet, Art, Macher, Bild, Livepicture, Showcase". Dem gegenüber konnte im nicht korrekten Pfad "Politikwissenschaft, öffentliches Recht" "öffentliches Recht" nicht als Teilgebiet von "Politikwissenschaften" gewertet werden. Innerhalb der bereinigten Ontologie konnten keine nicht korrekten Pfade identifiziert werden. So ist zum Beispiel eine "Universität" eine "Institution" (Pfad: "Institutionen, Universität"). Auch sind im Pfad "Data Mining, Examples" "Exampels" als Subbereich von "Data Mining" zu werten.



(a) Häufigkeitsverteilung der Begriffsbeschreibungen



(b) Wortkategorien (unbereinigte Ontologie)

(c) Wortkategorien (bereinigte Ontologie)

Abbildung 4.2: Statistiken von Begriffsbeschreibungen nach der Auswertung von Wissenskarten

4.1.3.2 Wissenskarten für einen geschlossenen Informationsraum

Für die zweite Studie wurden 26 Wissenskarten von zwölf Mitglieder des MARS Exploratory Media Labs ausgewertet, welche unter der Aufgabenstellung entstanden, damit einen geführten Rundgang zu erstellen. Bei der Auswertung dieser 26 Ablagestrukturen wurden insgesamt 69 Begriffsbeschreibungen für 134 Individuen generiert. Durch den Strukturvergleich ergaben sich 59 Rollen für die korrespondierenden Begriffsbeschreibungen. Das Filtern von Begriffsbeschreibungen mit einer Häufigkeit von eins mit ihren zugehörigen Rollen und Individuen führte zu einer bereinigten Ontologie mit 33 verbleibenden Begriffsbeschreibungen. Durch diese werden 82 Individuen beschrieben. Aufgrund des Filtervorgangs verblieben 32 der ursprünglich 147 Rollen in der bereinigten Ontologie. Zu Begriffsinklusionen konnten keine Untersuchungen durchgeführt werden, da Subsumtionsbeziehungen in dem verwendeten AWAKE-System vom Nutzer nicht definiert werden konnten.

Abbildung 4.2 (a) zeigt die absoluten Häufigkeiten der entstandenen Begriffsbeschreibungen. Die Begriffsbeschreibungen, welche am häufigsten vorkommen, sind "city" und "internet" mit jeweils vierfachem Vorkommen, "augmented reality" oder "theater", sowie "verteilte Systeme" oder "interaktive installation" für Beispiele von Wörtern mit dreifachen beziehungsweise doppeltem Vorkommen. In dieser Untersuchung zerfallen die 69 verwendeten Begrifflichkeiten in 63 Wörter für Themengebiete (zum Beispiel "Architecture" oder "Learning Platforms" und in sechs sonstige Wörter. Davon ist eines ungebräuchlich ("sofisuche") und fünf, wie "Überwachen" oder "Gewin-

ner" zu unspezifisch für eine klar umrissene Klassifikation. Dies entspricht nach Abbildung 4.2 (b) einem Anteil von 91% der Themenbegriffe beziehungsweise 9% der sonstigen Wörter. Die Themenbegriffe werden aus 26 Wörtern (38%) der deutschen Sprache und aus 37 (53%) Wörtern der englischen Sprache gebildet. Die Menge der sonstigen Begriffe besteht aus einem Prozent ungebräuchlicher Wörter und in acht Prozent unscharfer Begriffe. In Abbildung 4.2 (c) sind die absoluten und prozentualen Anteile der Wortkategorien innerhalb der bereinigten Ontologie gezeigt. Der Anteil der verbleibenden 30 Themenbegriffe beträgt immer noch 91%, welche von 10 Wörtern (30%) in deutscher Sprache und von 20 Wörtern (61%) in englischer Sprache gebildet werden. Der Bereich der drei (neun Prozent) sonstigen Wörter besteht aus einem ungebräuchlichen Wort (drei Prozent) und zwei unscharfen Wörtern (sechs Prozent).

Von den 134 entstandenen Individuen sind 69 multifacettiert und 65 monofacettiert von Begriffsbeschreibungen klassifiziert. Dies bedeutet einen Anteil von 52% der Individuen mit mehr als einer zugehörigen Begriffsbeschreibung und 48% von Individuen mit genau einer klassifizierenden Begriffsbeschreibung. Unter den Individuen mit den meisten Facetten ist eine Veröffentlichung über die Software "*machines will eat itself*"⁹ von Franz Alken, eines Systems, welches kommerzielle Webseiten mit fiktiven Nutzerprofilen versorgt. Damit soll der Wert der von diesen Webseiten erhobenen Nutzerdaten gemindert werden. Die Projektbeschreibung dieser Software befindet sich innerhalb der "Digital Sparks"-Sektion von "netzspannung.org" und ist einer der Gewinner dieser studentischen Ausschreibung im Jahre 2003. Die beschreibenden Begriffe zu dem Projekt sind "Arbeiten", "Sofisuche", "DS03 gewinner", "Digitale Gesellschaft", "Überwachen" und "Gewinner". Innerhalb der bereinigten Ontologie findet sich in der 82elementigen Individualmenge 26 (32%) multifacettierte und 56 (68%) monofacettierte Individuen. Eines der Individuen mit den meisten Facetten ist dabei eine Veröffentlichung über die interaktive Installation "*Coexistence*"¹⁰ der Künstler/innen Rebecca Allen, Eitan Mendelowitz und Damon Seeley. Diese wird durch die Begriffe "*Interactive Theatre*", "*Interaktive Installation*" und "*Theatre*" beschrieben.

Von den durch die 69 Begriffsbeschreibungen klassifizierten Individuen sind 133 korrekt und zwei falsch eingeordnet. Dies entspricht einer korrekten Klassifikationsquote von 99,3%. Diese Klassifikationsquote erhöht sich auf 100% innerhalb der bereinigten Ontologie, in der alle 82 Individuen korrekt eingeordnet sind.

Von den 147 generierten Rollen konnten alle (100%) als semantisch korrekt eingestuft werden. Zum Beispiel steht der Begriff "*Exhibition*" zu "*Information Spaces*", "*Internet*", "*Multimedia*" und "*Public Spaces*" in einer semantisch korrekten Verbindung. In der bereinigten Ontologie verbleiben aufgrund des Filtervorgangs 32 Rollen.

4.1.4 Diskussion

Die Ergebnisse der Evaluierung des Verfahrens zum kollaborativen Indexieren belegen dessen Anwendbarkeit zur Wissensakquise. Besonders erstaunt die Vielzahl der damit schon mit wenigen Nutzern generierten Begriffsbeschreibungen. Zwar besitzen die, mit den Ordnungsbezeichnungen vergebenen, Begrifflichkeiten keine durchwegs hohe Verständlichkeit, jedoch konnte durch die häufigkeitsbasierte Filterung der daraus entstandenen Begriffsbeschreibungen deren Qualität deutlich gesteigert werden. Im Vergleich zwischen offenen und geschlossenen Informationsräumen zeichnet sich eine Tendenz zur allgemein besseren Güte der vergebenen Begriffe im Hinblick auf die Verwendbarkeit zur Informationsrecherche bei geschlossenen Informationsräumen ab.

Am beeindruckendsten ist jedoch die äußerst hohe Güte der Begriffsassertionen. So sind rund 99% der Individuen korrekt klassifiziert. Dies stellt einen weiten Performanzvorsprung gegenüber rein maschinellen Klassifikationsverfahren, wie die Support-Vector-Maschine¹¹ (SVM) oder der Naive-Bayes-Klassifikator¹², dar. Deren Performanz wird in der Literatur mit maximal 66% angegeben.

⁹www.superbot.tk

¹⁰ALLEN/MENDELOWITZ/SEELEY: *Coexistence*, 2001

¹¹VAPNIK: *The Nature of Statistical Learning Theory*, 1995

¹²LEWIS: *Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval*, 1998

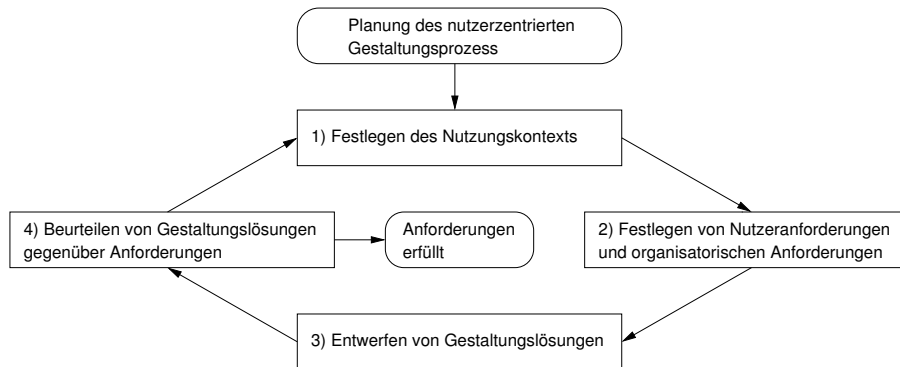


Abbildung 4.3: Benutzerorientierter Gestaltungsprozess nach ISO 13407

Ebenfalls belegt werden konnte die Annahme der inhaltlichen Verwandtschaft von Lesezeichenordnern, wenn diese gemeinsame Informationsobjekte beinhalten. Die daraus erzeugten Rollen von Begriffsbeschreibungen zeigen ebenfalls eine sehr hohe Güte. Bis auf wenige Ausnahmen sind diese semantisch korrekt. Ein weiterer Vorteil dieses Verfahrens stellt dessen Möglichkeit dar, hochwertige Begriffsinklusionen (Subsumtionsbeziehungen von Begriffsbeschreibungen) zu generieren. Auch hier kommen automatische Verfahren nicht an die vorliegende Performanzleistung heran.

4.2 Iterativer Gestaltungsprozess bei der Entwicklung des MatrixBrowsers

Der Strukturvisualisierung eines Informationsraums durch den MatrixBrowsers kommt in dem hier beschriebenen Ansatz besondere Bedeutung zu, da erst mit dieser die zugrunde liegende Ontologie navigierbar, explorierbar und abfragbar gemacht werden kann. Die Entwicklung des MatrixBrowsers erfolgte daher innerhalb eines iterativen und experimentell gestützten Vorgehens nach ISO 13407¹³. Der durch diese Norm definierte nutzerorientierte Gestaltungsprozess ist durch die folgenden Prinzipien gekennzeichnet:

1. Eine aktive Beteiligung der Nutzer und ein klares Verständnis von Nutzer- und Aufgabenanforderungen.
2. Eine geeignete Funktionsaufteilung zwischen Nutzern und Technik.
3. Die Iteration von Gestaltungslösungen.
4. Die multidisziplinäre Gestaltung.

Diese Prinzipien dienen dazu, die Nutzer, deren Aufgaben sowie die Benutzbarkeit des entstehenden Systems in Vordergrund zu stellen. Um dieses zu erreichen, werden die in Abbildung 4.3 dargestellten Aktivitäten identifiziert. Diese sind während der Entwicklung auszuführen. Im Einzelnen sind diese:

1. *Verstehen und Festlegen des Nutzungskontexts*: Die Nutzermerkmale, die Arbeitsaufgaben sowie die organisatorische und physische Umgebung bestimmen den Kontext, in dem das System verwendet wird. Es ist wichtig, die Einzelheiten dieses Kontextes zu verstehen und zu bestimmen, um frühzeitig Gestaltungsentscheidungen zu treffen und eine Grundlage für die Beurteilung zu liefern.

¹³INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: *ISO 13407: Human-centred Design Processes for Interactive Systems*, 1999

2. *Festlegen von Benutzeranforderungen und organisatorischen Anforderungen:* Neben den funktionellen und sonstigen Anforderungen an das Produkt oder System werden Benutzeranforderungen und organisatorische Anforderungen im Zusammenhang mit dem Nutzungskontext definiert.
3. *Entwerfen von Gestaltungslösungen:* Mögliche Gestaltungslösungen werden unter Berücksichtigung des entsprechenden Stands der Technik, der Erfahrungen und Kenntnisse der Teilnehmer und der Nutzungskontextanalyse entwickelt.
4. *Beurteilen von Gestaltungslösungen gegenüber Anforderungen:* Eine Beurteilung stellt einen bedeutenden Schritt bei der benutzerorientierten Gestaltung dar und sollte in jedem Stadium des Lebenszyklus des Systems stattfinden. Sie kann dazu benutzt werden, um Rückmeldung zugeben, die zur Verbesserung der Gestaltungslösung verwendet werden kann und inwieweit Nutzer- und Organisationsziele erreicht wurden.

Der Prozess wird als Iteration ausgeführt und endet, wenn der vierte Schritt mit einem zufriedenstellenden Ergebnis abgeschlossen ist.

Das Gestaltungskonzept des MatrixBrowsers wurde innerhalb dreier Iterationen entwickelt und zu Ende jeder Iteration evaluiert. Innerhalb des ersten Entwicklungszykluses wurde die Grundidee der Matrixvisualisierung im nicht interaktiven (statischen) Fall gegenüber konventionellen Netzrepräsentationen bewertet. Im nächsten Schritt wurden die interaktiven Grundkonzepte entwickelt, sowie die entsprechenden Darstellungs- und Interaktionskomponenten auf ihre Benutzbarkeit untersucht. Aufgrund der Ergebnisse dieser Studie wurde der MatrixBrowser weiter verfeinert und Gestaltungsalternativen erstellt. In der letzten Iteration wurden diese Alternativen im Fall interaktiver Nutzung mit einer strukturierten Netzdarstellung verglichen.

4.2.1 Erste Iteration

Der erste Schritt zur Entwicklung einer Visualisierung für ontologische Information bestand in der Evaluierung und der prototypischen Umsetzung der in Abschnitt 3.5.1, S. 99 beschriebenen Grundideen. Zunächst wurde eine einfache Matrixdarstellung implementiert, welche über wenig interaktive Fähigkeiten verfügte. Zudem war die horizontale Achsenbeschriftung senkrecht angeordnet. Zur Belegung der Achsen der Matrix mit Teilhierarchien eines darzustellenden Netzes wurden dem Nutzer zwei Auswahllisten für die jeweilige Achse zur Verfügung gestellt. Jede Auswahlliste zeigte dabei die Beschriftungen der Menge von Wurzelknoten der korrespondierenden Teilhierarchien. Bei Selektion eines Eintrages der Listen wurde die entsprechende Teilhierarchie auf einer der Achsen dargestellt. Zusätzlich wurden die unterschiedlichen, nicht hierarchischen Relationen des Ontologiegraphen anhand ihrer Typen sortiert und Beziehungen jeweils eines Typus als Zellen der Matrix auf verschiedene Registerkarten abgebildet. Der Reiter einer Registerkarte wurde mit dem Relationstyp beschriftet. Auf diese Weise konnte die anzuzeigende Menge einer Beziehungsart durch die Registerkartenreiter ausgewählt werden (vgl. Abbildung 4.4 (a)).

4.2.1.1 Vorgehensweise

Zur Evaluierung des Basiskonzeptes wurden zwei initiale Untersuchungen durchgeführt, in denen Suchzeiten und visuelle Suchstrategien in herkömmlichen Netzstrukturen mit denen der MatrixBrowser-Darstellung analysiert wurden. Dabei wurde die Organisationsstruktur der Fraunhofer Gesellschaft, wie sie auf ihrer Webseite "<http://www.fraunhofer.de>" zu finden ist, als Anwendungsszenario zugrunde gelegt. Einzelne Institute der Fraunhofer Gesellschaft sind darin zu Forschungsverbänden zusammen geschlossen, forschen in bestimmten Forschungsgebieten und sind auf verschiedene Standorte verteilt. Im Anwendungsszenario subsumiert die Fraunhofer Gesellschaft die Forschungsverbände, welche wiederum die Institute subsumieren. Auch die einzelnen Forschungsgebiete sind zu größeren thematischen Gebieten zusammengefasst. So beinhaltet die Informationstechnik Teilgebiete wie Softwaretechnik und virtuelle Realität. Die Städte,

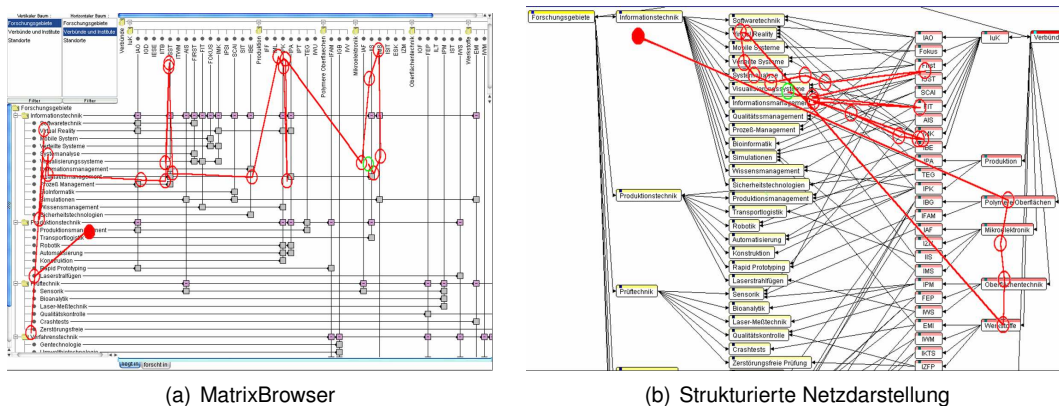


Abbildung 4.4: Visuelle Suchpfade von Probanden

in denen sich die verschiedenen Institute befinden, sind den entsprechenden Bundesländern innerhalb der Bundesrepublik Deutschland zugeordnet. Diese drei Hierarchien sind untereinander durch die Relationstypen "befindet sich in" und "forscht in" vernetzt.

In einer ersten Untersuchung wurde eine konventionelle Netzdarstellung ohne besondere visuelle Struktur mit einer korrespondierenden Matrix-Darstellung verglichen. Fünf Probanden, Mitarbeiter des Fraunhofer IAO, lösten dabei mehrere Suchaufgaben, wobei Suchzeiten, Fehler und Blickbewegungsspuren aufgenommen wurden. Ein Blickbewegungspfad besteht aus einer Blickfixation, wobei eine Blickfixation einen Punkt kennzeichnet, der mindestens 50 Millisekunden betrachtet wird. Drei der Probanden waren weiblich, zwei männlich. Zu suchen waren Begriffe und Relationen in einem statisch visualisierten Netzwerk und dem MatrixBrowser.

Da eine unstrukturierte Netzdarstellung bereits bei niedrigen Knoten- und Kantenzahlen praktisch nicht mehr überschaubar ist, wurde in einer zweiten Untersuchung innerhalb des selben Tests eine stark strukturierte Netzdarstellung mit expliziter räumlicher Zusammenfassung hierarchischer Teilstrukturen mit der Matrix-Darstellung verglichen.

4.2.1.2 Ergebnisse

Bei einem Graph mit 27 Knoten erforderte die Lösung der Aufgaben mit der konventionellen Netzdarstellung eine mittlere Suchzeit von 45 Sekunden, während mit der Matrix-Darstellung 18 Sekunden benötigt wurden. Die Suchzeiten sind damit durch den MatrixBrowser um 60% reduziert. Die mittlere Anzahl der Blickfixationen ergab sich bei der Netzdarstellung zu 51 und beim MatrixBrowser zu 24. Dies entspricht einer Reduktion der Blickfixationsanzahl um 53%. Die Blickbewegungsspur eines Probanden mit dem Matrixbrowser ist in 4.4 (a) dargestellt.

Im Vergleich mit der stark strukturierten Netzdarstellung zeigten sich leichte, allerdings nicht signifikante Vorteile des Matrix-Browsers, obwohl die Blickbewegungspfade wie in Abbildung 4.4 (a) und (b) deutlich gleichmäßiger verlaufen. Innerhalb der Netzdarstellung wurde eine durchschnittliche Suchzeit von 21 Sekunden gegenüber 19 Sekunden mit dem MatrixBrowser benötigt. Das entspricht einer Reduktion der Suchzeit um zehn Prozent. Die mittlere Anzahl der Blickfixation betrug 23 Fixationen bei der Netzrepräsentation und 21 Fixationen bei dem MatrixBrowser, was einer Verminderung der Blickfixationen um neun Prozent gleichkommt. Bei diesen Ergebnissen ist allerdings zu berücksichtigen, dass die interaktiven Fähigkeiten des MatrixBrowsers überhaupt noch nicht ausgenutzt wurden, um die Vergleichbarkeit der visuellen Suchaufgaben zu gewährleisten.

4.2.1.3 Diskussion

Obwohl die statische Matrixvisualisierung Performanzvorteile gegenüber anderen Netzdarstellungen zeigte, wurden intuitiv zwei Defizite des ersten Prototypen für den interaktiven Fall erkannt: Zum einen wurden die zwei Auswahllisten als nicht ausreichend erachtet, welche die Wurzelknoten der verfügbaren Teilhierarchien zur Belegung der Achsen zeigten. Sie vermitteln nicht in ausreichendem Maße eine Vorstellung, was sich inhaltlich in einer Teilhierarchie verbirgt und bieten nicht genügend interaktive Möglichkeiten, um die auf den Achsen dargestellte Knotenmenge flexibel auswählen zu können. Auch lässt sich damit nicht die volle Adjazenzmatrix darstellen. Zum anderen verfügte der erste Prototyp nicht über die Möglichkeit, Knotennachbarschaften und Relationen zwischen einem Knoten innerhalb einer der zwei gerade sichtbaren und einer n-ten nicht sichtbaren Hierarchie vollständig zu präsentieren. So steht im obigen Beispiel der Knoten "IAO" durch die Relationstypen "befindet sich in" mit dem Knoten "Stuttgart" und durch "forscht in" mit den einzelnen Forschungsgebieten in Beziehung. Jedoch ist immer nur ein Relationstyp innerhalb eines visuellen Zustands der Matrix sichtbar.

4.2.2 Zweite Iteration

Für den ständigen Überblick über die im Netz enthaltenen Teilhierarchien und Knotenmengen wurde im zweiten Entwicklungszyklus das in Kapitel 3.5.1 beschriebene Übersichtsfenster eingeführt. Dieses erlaubt die flexible Selektion von Teilstrukturen des Netzes indem sowohl Äste des Gesamtbaumes als auch Untermengen der Knoten via dem Mechanismus des Ziehen&Fallenlassens auf die einzelnen Achsen platziert werden können. Zusätzlich wurde die Matrixdarstellung um eine zusätzliche, alternative Visualisierung ergänzt, welche die Exploration der direkten Knotennachbarschaften erlaubt. Dabei wurde zuerst eine einfache Netzvisualisierung verwendet, bei der die Nachbarknoten zu einem gewählten Knoten gezeigt werden. Die Nachbarknoten sind dabei kreisförmig um einen Zentrumsknoten angeordnet, welcher den gewählten Knoten darstellt (vgl. Abbildung 4.5 (a)). Wird ein Nachbarknoten selektiert, so ergibt sich daraus der neue Zentrumsknoten. Falls ein Nachbarknoten nicht Teil der jeweils angezeigten Knotenmengen auf der Matrix ist, wird er besonders markiert. Gleichzeitig werden die Knotenmengen oder Hierarchien in der Matrix angezeigt, in denen der selektierte Knoten Mitglied ist.

4.2.2.1 Vorgehensweise

Zur Evaluierung der zusätzlichen interaktiven Ausgestaltung des MatrixBrowser-Konzepts wurde eine qualitative Benutzbarkeits- und Erlernbarkeitsanalyse mit weiteren fünf Probanden durchgeführt. Während des Tests lösten alle Probanden nacheinander Suchaufgaben auf Basis zweier unterschiedlich strukturierten Informationsnetzen sowohl mit dem MatrixBrowser als auch zum Vergleich mit einer Touchgraph¹⁴ genannten Software. Diese Software visualisiert Graphen als Netz, indem Knoten unter Zuhilfenahme virtueller Federn angeordnet werden. Die dadurch entstandene Repräsentation wird zudem auf eine hyperbolische Fläche projiziert. Beide Netze beinhalteten Information zu Restaurants, deren Standorten und Speisen beziehungsweise Spezialisierung (z.B. asiatische Küche). Das erste Netz bestand aus 89 Knoten, 3 Hierarchien und 849 Relationen. Das zweite Netz war aus 102 Knoten, 4 Hierarchien und 1146 Relationen aufgebaut. Für beide Datensätze ergaben sich bei verschiedenen Suchaufgaben unterschiedliche Navigationsschritte, die zur Lösung der Aufgaben nötig waren. In einem der Netze waren manche Aufgaben durch semantische Zuordnung lösbar, während sie im anderen Netz über rein syntaktische Suchvorgänge zu lösen waren. Die fünf Probanden waren ebenfalls Mitarbeiter des Fraunhofer IAO Stuttgart, die den Matrix Browser selbst noch nicht bedient hatten, wohl aber Kenntnis von ihm hatten. Keiner der Probanden kannte den Touchgraph, die zweite Software, die verwendet wurde. Zwei der Probanden waren weiblich, drei männlich.

¹⁴<http://www.touchgraph.com>

4.2.2.2 Ergebnisse

Als Hauptergebnisse dieser Evaluationsstudie lassen sich seitens der Probanden positive Rückmeldung im Hinblick auf die grundsätzliche Idee, die übersichtliche Matrixdarstellung, das intuitive Interaktionskonzept, sowie die leichte Erlernbarkeit festhalten. So wurde die Matrixdarstellung aufgrund ihrer geordneten Darstellung als nützlich für die Aufgabenbearbeitung bewertet, da komplexe Strukturen mit zahlreichen Knoten und Kanten im Vergleich zur Touchgraph-Visualisierung leichter erkennbar sind. Zwar wurde die Arbeit mit dem Touchgraph subjektiv als Spaß beurteilt, da die Darstellung von vielen Knoten aufgrund der Federwirkung fast nie zur Ruhe kommt, der MatrixBrowser aber wurde als ernsthafter, unkomplizierter und weniger belastend eingeschätzt. Dieses lässt sich auch mit der verkürzten Bearbeitungszeit der Aufgaben bei Verwendung des MatrixBrowsers belegen. Ebenfalls positiv wurden die dem "Windows Explorer" entlehnten Interaktionskonzepte, wie das Auf- und Zuklappen von Ästen eines Hierarchiebaumes und abgeleiteten Relationen, sowie die Ziehen&Fallenlassen Mechanismen zur Exploration gewünschter Teilstrukturen als vertraut beurteilt. Auch der Umgang mit den ToolTips, welche den semantischen Typ von Relationen darstellen, zeigte keine Probleme. Obwohl die Probanden einschätzten, nicht ohne fremde Hilfe mit dem Matrix Browser arbeiten zu können, hatten sie den Eindruck, dass die Einarbeitung wenig Zeit benötigt. So vertraten sie die Meinung, den Matrix-Browser nach der kurzen Einführung zu Beginn des Tests selbst bedienen zu können.

4.2.2.3 Diskussion

Neben dieser positiven Aufnahme wurde auch Optimierungsbedarf erkannt. Besitzt ein Knoten viele direkte Nachbarn (z.B. Restaurants in Berlin), ist die Netzdarstellung zur Darstellung dieser Nachbarn ungeeignet, da die Nachbarknotenbeschriftungen aufgrund ihrer Menge nicht mehr lesbar sind. In diesem Fall ist auch die ToolTip-Unterstützung wegen der zu engen Platzierung der dargestellten Knoten nur von eingeschränkter Hilfe. Ein weiterer Punkt betrifft die Lesbarkeit der horizontalen Matrixachse. Angesichts deren senkrechter Beschriftung fällt es schwer, bei normaler Kopfhaltung eine große Menge von Knoten zu erfassen. Zusätzlich wurde von vier der fünf Probanden der Wunsch nach einer Möglichkeit geäußert, bereits getätigte Interaktionsschritte wieder rückgängig zu machen und zu einem vorhergehenden interaktiven Zustand zurückkehren zu können.

4.2.3 Dritte Iteration

Aufgrund der vorangehenden Evaluationsergebnisse wurde im dritten Entwicklungszyklus die horizontale Achsenbeschriftung des MatrixBrowsers zur besseren Lesbarkeit um 45° gedreht. Zur besseren Darstellung von Knotennachbarschaften wurde die endgültige, in Abschnitt 3.5.1, S. 99 beschriebene, tabellenbasierte Visualisierungstechnik gewählt. Zusätzlich wurde die im selben Abschnitt erwähnte Interaktionshistorie ergänzt, so dass damit vorhergehende interaktive Zustände wieder hergestellt werden können.

Obwohl die Beschriftungen der Knoten auf der horizontalen Achse um 45° gedreht wurden, sollte noch eine weitere Gestaltungsvariante in Bezug auf bessere Lesbarkeit getestet werden. In dieser Variante wurde die komplette Matrixdarstellung um 45° gedreht, was eine horizontale Ausrichtung der Knotenbeschriftung zur Folge hat (siehe Abbildung 4.5 (b)). Auch wurde hier bewusst auf verschiedene Teile des MatrixBrowsers, wie die Nachbarschaftsvisualisierung oder das Übersichtsfenster verzichtet, um Unterschiede im Nutzungsverhalten der beiden Varianten besser herausarbeiten zu können. Die Matrix selbst, sowie die Bäume auf den Achsen, verhalten sich interaktiv jedoch wie beschrieben, mit der Ausnahme, dass die Achsen nicht mit Knotenmengen belegt werden können, sondern nur den einen Baum aus der Übersichtsdarstellung darstellen. Damit wird die volle Adjazenzmatrix des Netzes mit extrahierten Teilhierarchien gezeigt, deren sämtliche Querbeziehungen in den Zellen visualisiert werden.

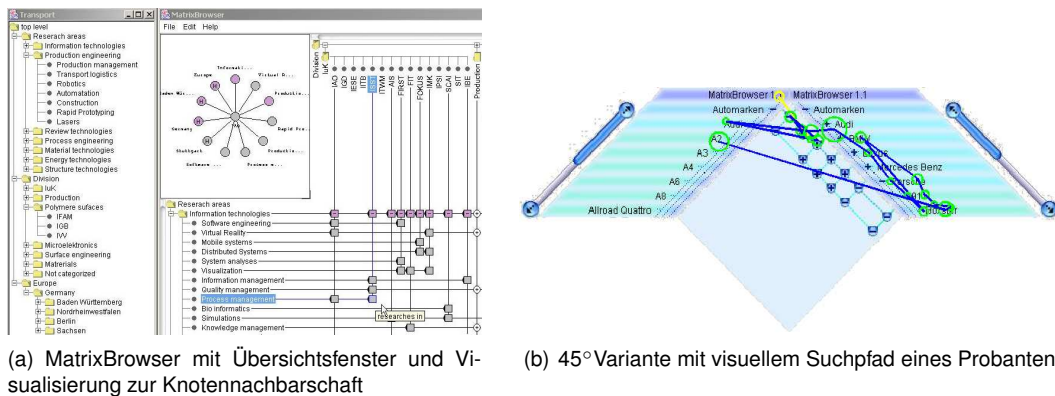


Abbildung 4.5: Gestaltungsvarianten des MatrixBrowsers

4.2.3.1 Vorgehensweise

In der dritten und letzten Untersuchung sollten im wesentlichen zwei Fragenkomplexe betrachtet werden. Einerseits sollte die allgemeine Benutzbarkeit des MatrixBrowser in diesem Entwicklungsstadium im Vergleich mit dessen, um 45° gedrehten, Variante und im Vergleich mit der Software *ST4* der Firma Schema GmbH¹⁵ betrachtet werden. Andererseits sollte die Performanz der drei Systeme im dynamischen Einsatz durch den Endbenutzer ermittelt und verglichen werden. Bei der Software *ST4* handelt es sich um ein kommerzielles Produkt zur technischen Dokumentation. Zur Visualisierung der Struktur einer solchen Dokumentation wird eine Technik eingesetzt, welche ebenfalls eine strukturierte Anordnung der Knoten mit räumlicher Zusammenfassung hierarchischer Teilstrukturen (z.B.: eine Kapitelstruktur) erlaubt. Eine Vorgängerversion von *ST4* kam für die erste Iteration der MatrixBrowser-Evaluierung zum Einsatz.

Als Probanden wurden vier Mitarbeiter des Fraunhofer IAO ausgewählt. Diese hatten zwar Kenntnis von dem MatrixBrowser, waren mit dessen Bedienung jedoch nicht vertraut. Keiner der Probanden kannte die Vergleichsprogramme - die 45° Matrix und *ST4*. Drei der Probanden waren weiblich, einer männlich. Da der MatrixBrowser die Kontrollanwendung war, haben ihn alle vier Probanden getestet. Die 45° Matrix und *ST4* wurde von jeweils zwei Probanden zusätzlich benutzt. Während des Tests lösten die Probanden Suchaufgaben aus der Domäne Autotypen mit Merkmalen wie Leistung, Farbe und Ähnlichkeit zu anderen Autos. Dabei wurden Suchzeiten, Fehler und Blickbewegungsspuren aufgezeichnet.

4.2.3.2 Ergebnisse

Nach einem Trainingsdurchlauf zeigte sich ein deutlicher Vorteil des MatrixBrowsers. Im Durchschnitt lösten die Probanden die gestellten Aufgaben mit dem MatrixBrowser in 25 Sekunden. Im Vergleich dazu benötigten sie 52 Sekunden mit der 45° Variante und 62 Sekunden mit dem *ST4* System. Im Mittel wurden in der Blickbewegungsanalyse mit dem MatrixBrowsers 16, mit der 45° Variante 20 und mit *ST4* 25 Fixationen zur Lösung einer Aufgabe benötigt.

4.2.3.3 Diskussion

Bemerkenswert ist hierbei, dass trotz der vermuteten besseren Lesbarkeit der 45° Variante des MatrixBrowsers die Performanzzeiten beim normalen MatrixBrowser um 52% und die Anzahl der Blickfixationen um 20% niedriger lagen. Dies begründet sich hauptsächlich in der vorhergehenden Extraktion hierarchischer Substrukturen und deren flexiblen Platzierung auf den Achsen, da hierdurch nicht die volle Adjazenzmatrix in die Exploration einbezogen werden musste. Insgesamt zeigen die bisherigen Ergebnisse, dass der MatrixBrowser auf Grund seiner regelmäßigen

¹⁵<http://www.schema.de>

Anordnungsstruktur effizienter als konventionelle Netzdarstellungen ist. Auch konnte die These begründet werden, dass gerade bei komplexen und großen Netzen durch die interaktiven Funktionen und die Aufbereitung des Netzes deutliche Vorteile gegenüber der herkömmlichen Netzdarstellung zu erzielen sind.

4.3 Evaluierung der Recherveschnittstelle

Nachdem die zentrale Visualisierungskomponente für die Darstellung ontologischer Informationsnetze einen befriedigenden Reifegrad erlangte, wurde auf Basis des MatrixBrowsers die Suchschnittstelle nach Kapitel 3.5, S. 98 prototypisch umgesetzt. Dieser Prototyp diente zur Evaluierung des Gesamtkonzepts aus Sicht der Nutzer.

4.3.1 Ziel und Fragestellung

Ziel der Studie war weniger die Identifikation von Benutzbarkeitsproblemen der verschiedenen Komponenten, als vielmehr inwieweit die in Kapitel 3.1, S. 68 formulierten Anforderungen erfüllt wurden. (Effektivität der Zugangs- und Abfragemechanismen, Förderung des Verständnis von Informationsraum und Informationsbedarf, inhaltliche Erfassung der Ergebnismenge einer Suchanfrage). Folgende weiteren Fragestellungen waren dabei ebenfalls von Interesse:

- Ist die Anwendungsdomäne richtig abgegrenzt?
- Wie hoch ist die Akzeptanz und wie wird die Nützlichkeit der drei Hauptfunktionalitäten von verschiedenen Nutzergruppen bewertet: Navigation im Informationsraum, Kontextualisierung der Suchergebnisse einer Stichwortsuche und präzise semantische Suche anhand ontologischer Metadaten?
- Kann der Prototyp auch von durchschnittlichen Nutzern zur Recherche verwendet werden oder handelt es sich dabei um ein reines Expertenwerkzeug?
- Wie hoch ist der allgemeine Lernaufwand?

4.3.2 Vorgehensweise

Die Untersuchung setzte sich aus prä- und postaktionalen Interviews, einem Fragebogen sowie eigens dafür zusammengestellten Aufgaben zusammen. Im präaktionalen Interview wurde der soziodemographische Hintergrund des Teilnehmers erfragt. Auch seine bisherige Erfahrung mit Visualisierungen und Werkzeugen zur Informationsbeschaffung wurde erhoben. Das postaktionale Interview fokussiert auf eine ausführliche Beurteilung (Nützlichkeit der Funktionalitäten, Lernaufwand, Anwendungsdomäne) des Prototyps aus Sicht der Nutzer. Ein standardisierter Fragebogen auf Grundlage der in DIN EN ISO 9241 "Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten"¹⁶ formulierten Dialoggrundsätze erlaubte die Bewertung der allgemeinen Benutzungsfreundlichkeit anhand der Dimensionen Aufgabenangemessenheit, Steuerbarkeit, Erwartungskonformität und Erlernbarkeit. Die anderen in der Norm vorgeschlagenen Dimensionen wurden aufgrund ihrer Irrelevanz für den vorliegenden Themenbereich bewusst vom Fragebogen ausgenommen.

Den Schwerpunkt des Tests bildete die Bearbeitung von typischen Aufgaben zur Informationsrecherche, wobei jeweils die Bearbeitungszeit gemessen wurde. Drei der Aufgaben entfielen auf die Stichwortsuche, in denen Themen recherchiert und Mehrdeutigkeiten eines Suchbegriffs aufgelöst werden mussten. Nach einer kurzen Einweisung erfolgte die Lösung der Aufgaben vergleichend sowohl mit dem Suchwerkzeug *Google*¹⁷ als auch mit einer Version des Prototypen, dessen Stichwortsuche ebenfalls auf *Google* beruhte. Als Ontologie kam eine semantisch

¹⁶INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: *ISO 9241: Ergonomic requirements for office work with visual display*, 1998

¹⁷<http://www.google.de>

aufbereitete Teilstruktur des *Open Directory Projects*¹⁸, eines frei verfügbaren Webkatalogs, zum Einsatz. Nachdem dieser Katalog sehr umfangreich ist, wurde die Teilstruktur auf den Themenbereich "Kunst"¹⁹ eingeschränkt. Fünf weitere Aufgaben entfielen auf die Bestimmung ontologischer Merkmale wie Attribute (z.B. Titel oder Autor) von Informationseinheiten und Relationen von Begriffsbeschreibungen. Eine Aufgabe befasste sich mit der Konstruktion semantischer Suchabfragen. Die Bearbeitung dieser Aufgaben erfolgte mit dem Prototypen auf Basis von *Google/Open Directory Project*- und *netzspannung.org*-Daten. Der Informationsraum von *netzspannung.org* hat im Gegensatz zu *Google/Open Directory Project* den Vorteil, dass er reicher an ontologischen Metadaten ist. Zur Vermeidung von Reihenfolgeeffekten wurde die Abfolge der zu bearbeitenden Aufgaben im Test vertauscht.

4.3.3 Ergebnisse

Insgesamt nahmen zehn Teilnehmer an der Evaluationsstudie teil. Fünf der Teilnehmer waren Mitarbeiter des Fraunhofer IAO und wurden als Expertennutzer eingestuft, welche mit der Thematik Ontologien und Semantik vertraut waren. Weitere fünf Teilnehmer kamen aus dem privaten Umfeld des Autors dieser Arbeit und repräsentierten eine Nutzergruppe, der zwar der Umgang mit dem Internet, nicht aber die speziellen Thematik der semantischen Suche bekannt war. Das Durchschnittsalter der Probanden betrug 28,6 Jahre. Die Geschlechterverteilung war mit vier weiblichen und sechs männlichen Teilnehmern in etwa ausgeglichen. Die erreichten Bildungsabschlüsse waren zu 70% Hochschulabschlüsse, die restlichen 30% hatten Abitur.

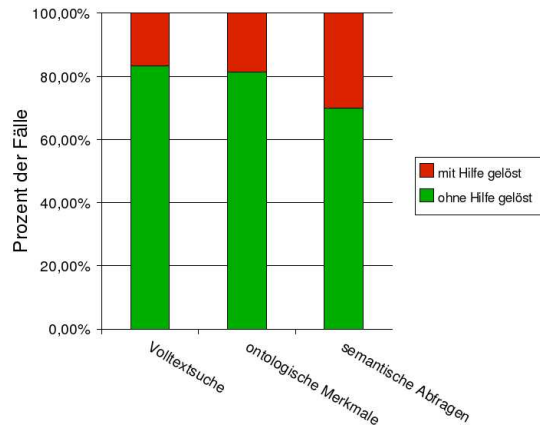
Die Quote von erfolgreich und ohne Hilfestellung gelösten Aufgaben ist ein Gradmesser der Schwierigkeit der Aufgaben. Im vorliegenden Fall bestehen die Schwierigkeiten im grundsätzlichen Verstehen des Gesamtkonzepts, dem Erfassen der semantischen Struktur der Benutzungsschnittstelle und die Intuitivität der Bedienbarkeit. Auch die inhaltliche Struktur des dargestellten Informationsraums fließt in die Bearbeitbarkeit der Aufgaben mit ein. Die Zielvorstellung für den Optimalfall wäre dementsprechend, dass alle Aufgaben ohne Hilfe des Moderators gelöst werden können.

Der prozentuale Anteil der Fälle für die verschiedenen Aufgabengruppen, bei denen die Aufgaben mit dem Prototypen ohne beziehungsweise mit Hilfestellung gelöst wurden, zeigt Abbildung 4.6 (a). Aus der Grafik wird deutlich, dass zwar alle Aufgaben von den Teilnehmern gelöst wurden, jedoch zum Teil Hilfestellung gegeben werden musste. So musste in zwei Fällen wiederholt werden, dass der Doppelklick auf ein Themengebiet (Knoten auf der Matrixachse) die zugehörigen Inhalte liefert und in drei Fällen nochmals erklärt werden, dass die Länge der Balken innerhalb eines Themengebietes mit der Anzahl der Treffer in diesem Themengebiet korreliert. Dies ergibt im Schnitt einen Anteil von 17% Hilfeleistung beim Bearbeiten der Stichwortsuchaufgaben mit dem Prototypen. Unter Verwendung der *Google*-Webseite konnten alle Aufgaben von allen Probanden ohne fremde Hilfe bearbeitet werden. Bei der Suche nach ontologischen Merkmalen musste in 19% der Fälle auf Hilfestellung seitens des Moderators zurückgegriffen werden. Die Hilfen, welche in diesen beiden Aufgabengruppen gegeben werden mussten, beziehen sich hauptsächlich auf die Begrifflichkeiten von Ontologien wie die Bedeutung von Rollen, Attributen und Begriffsbeschreibungen. Zur Konstruktion semantischer Suchaufgaben musste ebenfalls in 30% Hinweise zur Lösung gegeben werden. Hierbei war meist die Syntax von semantischer Abfragen nicht klar, das heißt in welcher Abfolge Begriffsbeschreibungen und Rollen anzuordnen sind. Innerhalb der Aufgabengruppen war ein Lerneffekt festzustellen, da meist nur bei der ersten Aufgabe innerhalb einer Gruppe Hilfestellung gegeben werden musste. Die darauf Folgenden konnten dann selbständig gelöst werden. Wie zu erwarten war, mussten in der Laiengruppe öfters Hinweise gegeben werden als in der Expertengruppe. Der Unterschied ist allerdings nicht signifikant.

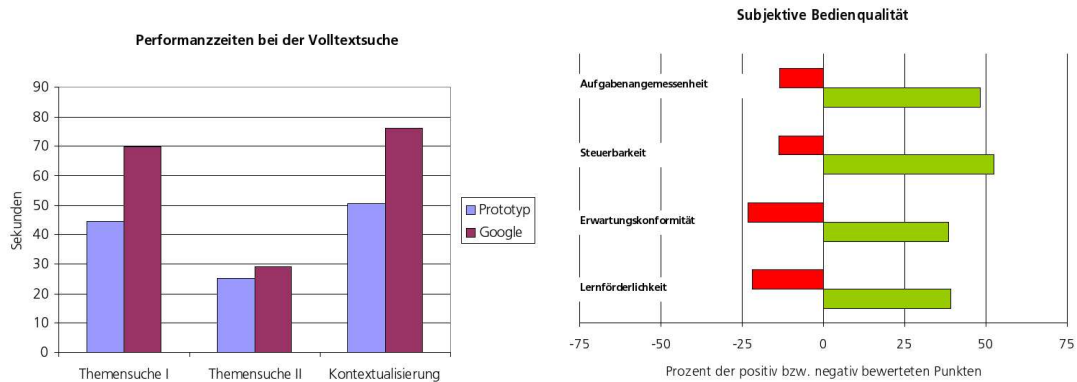
Abbildung 4.6 (b) zeigt die durchschnittliche Bearbeitungszeit der Stichwortsuchaufgaben mit dem Prototypen und mit *Google*. Dabei zeigen sich teilweise ausgeprägte Performanzvorteile bei dem Prototypen. In der ersten Aufgabe wurden mit dem Prototypen im Mittel 44,7 Sekun-

¹⁸<http://www.dmoz.org>

¹⁹<http://www.dmoz.org/Arts/>



(a) Aufgabenlösungen nach Aufgabengruppen



(b) Performanzzeiten Volltextsuche

(c) Subjektive Bedienqualität des Prototypen nach ISO 9241

Abbildung 4.6: Ergebnisse der Evaluationsstudie der Recherveschnittstelle

den und mit *Google* 69,7 Sekunden benötigt. Dies entspricht einem Performanzvorteil 35,7%. Dieser Effekt wurde durch Lösung der Aufgabe innerhalb des Prototypen mit einem Doppelklick auf die gesuchte Kategorie erzielt, während bei *Google* eine zielführende Kombination aus zwei Suchbegriffen gefunden werden musste. Ein weniger ausgeprägter Vorteil (12,6%) ergab sich bei der zweiten Themenrecherche mit einem Durchschnitt von 25,4 Sekunden beim Prototypen gegenüber 29,4 Sekunden mit *Google*. Hierbei genügte bei *Google* allerdings schon die Eingabe von nur einem Suchwort. Ein deutlicher Vorteil des Prototypen (33,6%) ergab sich bei einer Aufgabe zur Auflösung von Mehrdeutigkeiten (Kontextualisierungsfunktion). Es wurden im Mittel 50,6 Sekunden mit dem Prototypen und 76,2 Sekunden mit *Google* zur Lösung dieser Aufgabe benötigt. Auch hier ist der Unterschied in den Performanzzeiten zwischen der Laien- und Expertengruppe nicht signifikant. So wurden manche Aufgaben mal von der einen Gruppe und mal von der anderen Gruppe schneller bearbeitet. Erstaunlich ist, dass die Aufgaben auch von der Laiengruppe erfolgreich bearbeitet werden konnten, obwohl die Begrifflichkeiten aus der Terminologie von Ontologien dieser zum Teil unbekannt war.

Die Ergebnisse des Fragebogens zur allgemeinen Benutzungsfreundlichkeit des Prototypen auf Grundlage der in DIN EN ISO 9241 ist in Abbildung 4.6 (c) dargestellt. Hier wird der Anteil der vergebenen Punkte einer bestimmten Dimension angegeben, die entweder negativ (Werte von -2 oder -1) oder positiv (Werte von +1 oder +2) bewertet wurden. Der fehlende Anteil zu 100% besteht aus Punkten, die neutral (0) bewertet wurden. Die roten Balken können als relative

Schwächen und die grünen Balken als relative Stärken des Produkts bewertet werden. Wie aus Abbildung 4.6 (c) ersichtlich, fällt die Bewertung der subjektiven Bedienqualität sehr positiv aus. Besonders gut schneiden die Dimensionen Aufgabenangemessenheit mit 48% und Steuerbarkeit mit 52% positiv vergebenen Punkten ab. Diese sind für die Beantwortung der Frage nach der Nützlichkeit der Hauptfunktionalitäten des Recherchekonzepts von besonderer Bedeutung, da sie diese belegen. Die überwiegend positiv bewerteten Dimensionen Erwartungskonformität (38%) und Erlernbarkeit (39%) zeigen die Intuitivität der Benutzungsoberfläche. Dem gegenüber stehen allerdings auch deren kritische Befunde (Erwartungskonformität 23% und Erlernbarkeit 22%). Dies begründet sich nach Meinung des Autors in der Komplexität des Ansatzes und der Bedienoberfläche, sowie der zum Teil neuen und daher ungewohnten Interaktionskonzepte.

Diese Annahme wird in den Äußerungen der Probanden im postaktionalen Interview bestätigt. So wurde auf die Frage, wie gut die Probanden bei der Bearbeitung der Aufgaben zurecht gekommen seien, besonders von der Laiengruppe, die Vielzahl und Ungewöhnlichkeit der Funktionen des Prototyps angemerkt. Von beiden Gruppen wurde eine fundierte Einführung gefordert. Wurde jedoch Sinn und Zweck des Konzepts von den Probanden erkannt, bestätigten alle Teilnehmer die Nützlichkeit und Bedienbarkeit der Einzelfunktionen, wie Stöbern in ontologischen Merkmalen und Inhalten, Kontextualisierung von Suchergebnissen und semantischen Abfragen. Dabei wurde von 44% der Aussagen eine sehr gute, von 36% eine gute, von 16% eine mittelmäßige und von sechs Prozent eine schwere Bedienbarkeit festgestellt. Es erstaunt, dass auch hier die 20% der mittelmäßigen bis schweren Bedienbarkeitsbeurteilungen gleichermaßen von der Laien- und der Expertengruppe kamen.

Auf die Frage nach möglichen Anwendungsfällen gaben nur 20% der Probanden an, es handle sich dabei um ein Expertenwerkzeug. Weiter 40% gaben die allgemeine Recherche, wie die Suche nach Literatur, in speziellen Informationsräumen (Bibliotheken, Intranets, Datenbanken) an. Die restlichen 40% entfielen auf die gezielte Informationssuche in allgemeinen Informationsräumen. Ein Vorschlag war hierbei, den Prototypen als alternatives Suchwerkzeug für *Google* anzubieten.

Im Schnitt wurde die Einarbeitungszeit auf ungefähr zwei Stunden eingeschätzt. In dieser subjektiven Einschätzung zeigen sich wiederum Unterschiede zwischen der Laien- und Expertengruppe. In der Laiengruppe betrug die geschätzte mittlere Einarbeitungszeit rund drei Stunden, während die Expertengruppe diese auf rund 1 Stunde schätzte.

4.3.4 Diskussion

Insgesamt zeigte die Evaluationsstudie der Recherveschnittstelle deren überaus positive Aufnahme. Ebenso konnte durch sie die Erfüllung der formulierten Anforderungen an den vorgestellten Ansatz bestätigt werden. Gerade die Nützlichkeit der drei Hauptmerkmale des Ansatzes wie die Navigation und Exploration von Informationsräumen auf Basis deren ontologischer Abstraktion, die Möglichkeiten einer Stichwortbasierten- und semantischen Suche, sowie deren Integration in einer einheitlichen Benutzungsschnittstelle konnte belegt werden. Die gewählten Einsatzszenarien zur allgemeinen Recherche in spezialisierten Informationsräumen wie Intranets oder Wissensdatenbanken stellte sich als richtige Abgrenzung der Anwendungsdomäne heraus. Zusätzlich wurde von den Teilnehmern der Studie eine weitere Anwendbarkeit in allgemeinen Informationsräumen wie dem gesamten Internet angenommen.

Auf die Frage nach der Nutzergruppe konnte gezeigt werden, dass die prototypische Umsetzung der Recherveschnittstelle gleichermaßen von Experten und durchschnittlichen Nutzern zu bedienen war, obwohl die letztere Gruppe von der Komplexität und Vielzahl der interaktiven Möglichkeiten zunächst verunsichert wurde. Bei einer wirtschaftlichen Verwertung des Ansatzes besteht daher die Notwendigkeit, Nutzer gründlich in die Basiskonzepte und Bedienung einzuführen. Jedoch scheint der Einlernaufwand relativ gering zu sein, nachdem dieser im Mittel auf ungefähr zwei Stunden geschätzt wurde. Dies zeigt die Anwendbarkeit des Ansatzes in Recherche- und Suchszenarien des täglichen Lebens.

Abschließend ist zu bemerken, dass sich zwar Performanzvorteile gegenüber der Stichwortsuche mit *Google* für die in der Evaluationsstudie gewählten Suchaufgaben zeigen, diese decken

jedoch nicht alle Anwendungsfälle für *Google* ab. Damit kann kein endgültiges Fazit im Vergleich der beiden Recherchemöglichkeiten gebildet werden.

5 Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde ein nutzerorientierter und ganzheitlicher Ansatz für ein interaktives Computersystem entwickelt und evaluiert, welcher eine semantische Erschließ- und Recherchierbarkeit von großen und vernetzten Informationsräumen unter Einbeziehung der menschlichen Nutzung und Interpretation ermöglicht. Unter dem Begriff Informationsraum werden außer Dokumentkollektionen auch sonstige Ressourcen wie Dienste und Produktdaten zusammengefasst. Den Anwendungskontext stellen Intranets und komplexe Portalseiten, sowie spezialisierte Dokumentkollektionen dar, welche von Expertengruppen zur Dokumentation gemeinsamer Erfahrungen erstellt wurden. In diesen Anwendungsgebieten ist ein effektiver und nutzerfreundlicher Informationszugriff essentiell.

Ausgehend von der Betrachtung und Diskussion bestehender Methoden des klassischen und wissensbasierten Information Retrievals, sowie der Analyse von Anforderungen an ein ganzheitliches Recherchesystem wurde ein integrierter Ansatz vorgestellt und begründet. Ausgangspunkt war dabei die Annahme, dass in Situationen der Informationssuche, ein Kommunikationsprozess zwischen dem menschlichen Nutzer und dem benutzten Werkzeug innerhalb eines soziotechnischen Systems statt findet, welcher es erlaubt, aus der Vielzahl vorhandener Information die Passende zu selektieren und zu übertragen. Erst die Betrachtung als geschlossenes soziotechnisches System erlaubt die Vermittlung von Information und, durch deren Verarbeitung, eine Wissensbildung auf Seiten der Kommunikationspartner. Die Grundlage dazu stellt die Informations- und Kommunikationstheorie von Niklas Luhmann dar.

Die in diesem Kommunikationsprozess entstandenen oder auch bereits vorhandenen Artefakte, wie explizierte Präferenzen und persönliche Klassifikationssysteme, die sich in Lesezeichen zur Merkhilfe oder in Ordnerstrukturen manifestieren, werden im gewählten Ansatz zur Wissensbildung auf Seiten der maschinellen Verarbeitung genutzt und in einer Ontologie formalisiert. Ferner bezieht der Ansatz zusätzlich zu der herkömmlichen Volltextsuche mit Stichwörtern die tatsächliche semantische Suche auf Grundlage einer ontologischen Wissensbasis und eine geeigneten Nutzerschnittstelle zum Dialog mit dem Recherchesystem ein. Eine Ontologie als Modellierungsformalismus ermöglicht zudem die einfache Verknüpfungsmöglichkeit dieser mit nicht textuellen, strukturierten Informationsquellen, wie zum Beispiel Datenbankschemata. Zusätzliche Metadaten wie Autor, Erscheinungsjahr oder Bearbeiter lassen sich ebenfalls einbetten. Die systemische Wissensbasis hat eine Doppelrolle inne: Zum einen dient sie als semantische Abstraktion und Verdichtung eines Informationsraums als Grundlage zur interaktiven Visualisierung desselben. Zum anderen wird sie direkt zur Informationsrecherche genutzt und nicht nur als Thesaurus zur Anfrageerweiterung einer Stichwortsuche, wie bei vielen als semantisch bezeichneten Suchmaschinen. Zur Berücksichtigung der bei Kommunikationsprozessen immanenten Unsicherheiten und Vagheiten wurde die Ontologie mit Wahrscheinlichkeiten erweitert.

Bei der Entwicklung des gewählten Ansatzes galt es zunächst die Anforderungen an die Wissensbasis zu ermitteln. Aufgrund deren Anforderungen wurde eine standardisierte Spezifikationsprache ausgewählt, entsprechend erweitert, sowie ein grundsätzliches semantisches Modell im Sinne einer Minimalstruktur abgeleitet. Die Elemente der Minimalstruktur dienen dabei als Basis zur reichhaltigen semantischen Ausgestaltung des Informationsraums und als notwendige Grundlage für die Abbildung der Wissensbasis auf eine interaktive Visualisierung und Informationsrecherche. Zur Abfrage dient ein integriertes Suchmodell, welches die Volltextsuche und die semantische Suche vereint. Die enge Kopplung dieses Suchmodells mit der ontologischen Wissensbasis und der Nutzerschnittstelle ermöglicht die Bearbeitung von Suchaufgaben von der allgemeinen Exploration über die ungenaue Suche mit Stichwörtern bis zur präzisen Abfrage anhand bedeutungsvoller Merkmale.

Ferner wurde eine kollaborative Methode zur systemischen Wissensakquise vorgestellt, welche es erlaubt, nutzererstellte Informationsablagestrukturen zu analysieren und in die Wissensbasis einzuarbeiten. Mit dieser Methode wird eine Ontologie mit der Fähigkeit des Lernens ausgestattet und kann so in ihrem Lebenszyklus verfeinert, sowie an aktuelle Nutzerbedürfnisse angepasst werden.

Schließlich wurde die Nutzerschnittstelle beschrieben, mit deren Hilfe Ontologien visuell prä-

sentiert und durch eine graphische Abfragesprache interaktiv abgefragt werden kann. Die Grundidee hierzu ist die Verwendung einer hochinteraktiven Darstellung einer Adjazenzmatrix des zugrunde liegenden Ontologiegraphen. Durch die visuelle Konstruktion von semantischen Abfragen entfällt die Erlernung einer logischen Abfragesprache. Ein weiterer Aspekt dieser Nutzerschnittstelle ist die thematische Kontextualisierung von Stichwortsuchergebnissen in den damit verbundenen Ausschnitten der Ontologie, welche im aktuellen Recherchekontext relevant sind. Damit kann die Suche weiter verfeinert und zielgerichtet erfolgen.

Wie aus dieser Zusammenfassung der Ergebnisse hervorgeht, konnten die Hauptziele dieser Arbeit erreicht werden. Diese wurden in Abschnitt 1.2, S. 18 des ersten Kapitels formuliert. Der gewählte Ansatz wurde in Kapitel 3.2.1, S. 71 unter Einbeziehung alternativer Lösungswege ausführlich begründet. Im Rahmen der umfangreichen Evaluation konnte ebenfalls gezeigt werden, dass die genannten Anforderungen (vgl. Abschnitt 3.1, S. 68) an einen ganzheitlichen und bedarfsgerechten Erschließungs- und Rechercheansatz erfüllt wurden. So zeigen die entwickelten Zugangs- und Abfragemechanismen eine hohe Effektivität. Es konnten Leistungsvorteile gegenüber der herkömmlichen Volltextsuche nachgewiesen werden. Ebenfalls bestätigt wurde in der Evaluation eine Förderung des Verständnisses von Informationsraum und Informationsbedarf, sowie der inhaltlichen Erfassung der Ergebnismenge einer Suchanfrage mit Hilfe der eingesetzten Visualisierungs- und Interaktionstechniken. Zusätzlich erlaubt das vorgestellte Verfahren zur kollaborativen Wissensakquise die erfolgreiche Inkorporation von persönlichen Sichtweisen und Interpretationen bei der Nutzung von recherchierter beziehungsweise aufgefundener Information. Die detaillierte Evaluation und deren Ergebnisse wurde in Kapitel 4, S. 109 vorgestellt und diskutiert.

Ein zentraler Aspekt des vorgestellten Ansatzes ist die Visualisierung der semantischen Struktur eines Informationsraums mit Hilfe der Matrixdarstellung. Zwar wurde diese innerhalb eines iterativen und experimentell gestützten Vorgehens entwickelt, es besteht jedoch weiterhin Optimierungsbedarf. So sind Ontologien für den Einsatz im täglichen Leben sehr umfangreich. Die in der abschließenden Evaluationsstudie verwendete Ontologie hatte schon circa 4000 Knoten, stellte aber nur einen kleinen Ausschnitt des *Open Directory Projects* dar, welches ungefähr 80000 Knoten besitzt. Der Aufwand zum Blättern innerhalb der auf den Achsen der Matrix gezeigten Bäume wird dadurch sehr hoch. In dieser Richtung besteht daher noch Forschungsbedarf zur Reduktion desselben.

Ferner erfordert der Umgang mit der Nutzerschnittstelle aufgrund ihrer vielzähligen Präsentations- und Interaktionsmöglichkeiten eine gewisse Einarbeitungszeit. Durch weitere Forschung könnte diese verkürzt werden. Dies könnte zum Beispiel durch Reduktion der Komplexität der Nutzerschnittstelle oder deren Adaption an den Kenntnisstand des jeweiligen Nutzers geschehen. Auch weitere Interaktionshilfen, wie ein größerer Umfang der beschriebenen ToolTips, wären denkbar.

Ein weiterer Punkt betrifft die softwaretechnische Umsetzung des Ansatzes. Wie bereits gesagt, können Ontologien sehr umfangreich werden. Bei deren Umfang ist die Einsetzbarkeit von herkömmlichen Deduktionsmaschinen nicht mehr gewährleistet, da diese zur automatischen Schlussfolgerung sehr viel Ressourcen wie Speicher und Rechenleistung benötigen. Schon bei 4000 Knoten musste die Deduktionskomponente bei der eingesetzten Ontologieverwaltungssoftware *JENA*¹ deaktiviert werden, da zwei Gigabyte Hauptspeicher nicht mehr ausreichend war. Gerade auf Seiten der Werkzeugunterstützung des *Semantischen Netzes* besteht noch erheblicher Forschungs- und Entwicklungsbedarf.

¹MCBRIDE: *Jena: A Semantic Web Toolkit*, 2002

Literaturverzeichnis

- Agrawal, Rakesh/Imielinski, Tomasz/Swami, Arun:** Mining Association Rules between Sets of Items in large Databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data. ACM Press, 1993, ISBN 0-89791-592-5, 207-216
- Alber, J.:** Versorgungsklassen im Wohlfahrtsstaat. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 36 1984, 225-251
- Allen, James:** Natural Language Understanding. 2. Auflage. Redwood City, California: Cummings Publishing, 1995, ISBN 0-8053-0334-0
- Allen, Rebecca/Mendelowitz, Eitan/Seeley, Damon:** Coexistence. In: Proceedings of cast01: living in mixed realities. Fraunhofer IMK 2001
- Baeza-Yates, Ricardo A./Ribeiro-Neto, Berthier A.:** Modern Information Retrieval. Addison-Wesley, 1999, ISBN 0-201-39829-X
- Barth, R.:** 5000 Jahre Bibliotheken - eine Geschichte ihrer Benutzer, Bestände und Architektur. Skriptum zur Vorlesung. WWW, 1996 (URL: <http://www.stub.unibe.ch/stub/vor196/>)
- Bechhofer, Sean et al.:** OilEd: a Reason-able Ontology Editor for the Semantic Web. In: Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence. Vienna: Springer-Verlag, September 2001, Lecture Notes in Computer Science 2174, 396-408
- Bekavac, Bernard:** Skript zum Kurs Information Retrieval. Wintersemester 01/02. WWW, 2001, Lehrstuhl Prof. Kuhlen, Uni Konstanz (URL: http://www.inf-wiss.uni-konstanz.de/CURR/winter0102/IR/ir_script_ws01.pdf)
- Belkin, Nicholas J./Croft, Bruce W.:** Retrieval Techniques. In: Annual Review of Information Science and Technology (ARIST), 22 1987, 112
- Belkin, Nicolas J.:** Anomalous States of Knowledge as a Basis for Information Retrieval. In: Canadian Journal of Information Science, 5 1980, 133-143
- Berners-Lee, T.:** Semantic Web Roadmap. World Wide Web Consortium, 1998 – Technischer Bericht (URL: <http://www.w3.org/DesignIssues/Semantic>)
- Bertin, J.:** Graphics and Graphic Information-Processing. Berlin: Walter de Gruyter Co., 1981
- Blum, Rudolf:** Kallimachos und die Literaturverzeichnung bei den Griechen. Untersuchung zur Geschichte der Bibliographie. In: Archiv für Geschichte des Buchwesens. Saur Verlag, 1977, 18, ISBN 3-598-24814-8, 1-330
- Boardman, Richard:** Multiple Hierachies in User Workspace. In: Extended Abstracts of CHI Conference on Human Factors in Computing Systems. Seattle: ACM Press, 2001
- Borgatti, Stephen:** Elicitation techniques for cultural domain analysis. In: *Schensul, J./LeCompte, M. (Hrsg.): The Ethnographer's Toolkit. Band 3, Walnut Creek: Altamira Press, 1998*
- Borgida, A. et al.:** CLASSIC: A Structural Data Model for Objects. In: Proceedings of ACM SIGMOD International Conference on Management of Data. June 1989, 59-67
- Brachman, R.J./Schmolze, J.G.:** An Overview of the KL-ONE Knowledge Representation System. In: Cognitive Science, 2 1985, Nr. 9, 171-216
- Braschler, Martin/Ripplinger, Bärbel:** How Effective is Stemming and Decompounding for German Text Retrieval? In: Information Retrieval, 7 2004, Nr. 3-4, 291-316

- Brickley, D./Guha, R. V.:** RDF Vocabulary Description Language 1.0: RDF Schema. W3C, Februar 2003 – Working Draft
- Brockhaus:** Der Brockhaus multimedial 2000 Premium. Mannheim: Verlag Bibliographisches Institut und F.A. Brockhaus AG, 2000
- Bullinger, Hans-Jörg/Müller, Martin/Ribas, Miguel:** Wissensbasierte Informationssysteme - Enabler für Wissensmanagement. Stuttgart: IAO, 1999
- Büning, Hans Kleine/Lettmann, Theodor:** Logik und Regelverarbeitung, Skriptum zur Vorlesung Wissensbasierte Systeme. Institut für Informatik, Universität Paderborn, 2002 (URL: <http://www.uni-paderborn.de/cs/ag-klbue/de/courses/ws02/logicrules02/wbs-script-frame.pdf>)
- Capurro, Rafael:** Skript zur Vorlesung Wissensmanagement. WWW, 2001 (URL: <http://www.capurro.de/WM/bausteine.htm>)
- Card, S.K./Mackinley, J. D./Shneiderman, B.:** Readings in Information Visualization: Using Vision to Think. San Francisco: Morgan Kaufmann, 1999
- Catarci, Tiziana et al.:** An Ontology Based Visual Tool for Query Formulation Support. In: Proceedings of OTM 2003, On The Move to Meaningful Internet Systems 2003. Springer-Verlag Heidelberg, Oktober 2003, ISBN 3-540-20494-6
- Chaffey, Dave:** E-Business and E-Commerce Management: Strategy, Implementation and Practice. Harlow, UK: Pearson Education, August 2001
- Cleverdon, Cyril W.:** Progress in Documentation. Evaluation of Information Retrieval Systems. In: Journal of Documentation, 1 1970, Nr. 26, 55-67
- Clocksin, W. F./Mellish, C. S.:** Programming in Prolog. 3. Auflage. Berlin, Heidelberg: Springer, 1987
- Cohen, P.R./Kjeldsen, R.:** Information Retrieval by Constraint Spreading Activation in Semantic Networks. In: Information Processing and Management, 23 1987, Nr. 4, 255-268
- Crestani, F.:** Application of Spreading Activation Techniques in Information Retrieval. In: Artificial Intelligence Review, 11 1997, Nr. 6, 453-82
- Croft, Bruce W.:** Approaches to Intelligent Information Retrieval. In: Information Processing and Management, 1987, 249-254
- Croft, Bruce W./Harper, D.J.:** Using Probabilistic Models of Document Retrieval without Relevance Information. In: Journal of Documentation, 35 1979, Nr. 4, 285-295
- Cutting, Douglass R. et al.:** Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In: Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Kopenhagen, 1992, 318-329
- Delgrande, J.P./Mylopoulos, J.:** Knowledge Representation: Features of Knowledge. In: *Bibel, W./Jorrand, P. (Hrsg.): Fundamentals of Artificial Intelligence: An Advanced Course.* Berlin: Springer, 1986
- Deutsches Institut für Normung:** DIN 19226: Regelungstechnik und Steuerungstechnik, Begriffe und Benennungen. 1968
- Deutsches Institut für Normung:** DIN 1463-1: Erstellung und Weiterentwicklung von Thesauri; Einsprachige Thesauri. 1987

- Dublin Core Metadata Initiative:** Dublin Core Metadata Element Set, Version 1.1: Reference Description. Juni, 2004 (URL: <http://dublincore.org/documents/dcmi-terms/>)
- Esswein, Werner:** Das Rollenmodell der Organisation: Die Berücksichtigung aufbauorganisatorischer Regelungen in Unternehmensmodellen. In: Wirtschaftsinformatik, 35 1993, Nr. 6, 551–561
- Farquhar, A./Fikes, R./Rice, J.:** The Ontolingua Server: A Tool for collaborative Ontology Construction. Stanford, 1996 (KSL 96-26). – Technischer Bericht
- Faure, David/Nédellec, Claire/Rouveirol, Céline:** Acquisition of Semantic Knowledge using Machine Learning Methods: The System ASIUM. Laboratoire de Recherche en Informatique, Inference and Learning Group, Université Paris, 1998 (ICS-TR-88-16). – Technical report (URL: <http://www.lri.fr/ia/articles/faure/1998/faure98b.ps>)
- Feigenbaum, E.A./Barr, E./Cohen, P.:** The Handbook of Artificial Intelligence. Band 1-3, Los Altos, California, USA: Kaufman Inc, 1981
- Feldman, Susan:** NLP meets the Jabberwocky - Natural Language Processing in Information Retrieval. In: Online, 23 1999, Nr. 3, 62–72
- Fensel, Dieter:** Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Berlin, Heidelberg, New York: Springer-Verlag, 2000
- Fensel, Dieter et al.:** Ontobroker: The Very High Idea. In: Proceedings of the Eleventh International Florida Artificial Intelligence Research Society Conference. AAAI Press, 1998, ISBN 1–57735–051–0, 131–135
- Fensel, Dieter/Groenboom, Rix:** Specifying Knowledge-Based Systems with reusable Components. In: Proceedings of the 9th International Conference on Software Engineering and Knowledge Engineering, SEKE'97. Madrid, Spain, 1997
- Fensel, Dieter et al.:** OIL in a Nutshell. In: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management. Springer-Verlag, 2000, ISBN 3–540–41119–4, 1–16
- Ferber, Reginald:** Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. Heidelberg: dpunkt.verlag, 2003
- Fernandez-Lopez, M./Gomez-Perez, A./Juristo, N.:** METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In: AAAI-97 Spring Symposium on Ontological Engineering. Stanford, USA: AAAI Press, März 1997
- Fleischmann, Monika/Strauss, Wolfgang/al, Jasminko Novak. et:** netzspannung.org: an Internet Media Lab for Knowledge Discovery in Mixed Realities. In: Proceedings of cast01: living in mixed realities. Fraunhofer IMK 2001, ISSN 1618–1379
- Fowler, Richard H./Fowler, Wendy A. L./Wilson, Bradley A.:** Integrating Query Thesaurus and Documents through a common visual Representation. In: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, 1991, 142–151
- Fuhr, Norbert:** Probabilistic Datalog - A Logic for powerful Retrieval Methods. In: Proceedings of the 18th annual international ACM SIGIR. ACM Press, 1995, ISBN 0–89791–714–6, 282–290
- Fuhr, Norbert:** Gesellschaft für Informatik: Ziele und Aufgaben der Fachgruppe „Information Retrieval“. WWW, 1996 (URL: <http://www.is.informatik.uni-duisburg.de/fgir/mitgliedschaft/brochure2.html>)

- Fuhr, Norbert:** Information Retrieval Skriptum zur Vorlesung im SS04. WWW, Februar 2004 (URL: http://www.is.informatik.uni-duisburg.de/teaching/lectures/ir_ss04/folien/irskall.pdf)
- Furnas, G.W.:** Generalized Fisheye Views. In: Proceedings of ACM CHI'86,. ACM ACM Press, 1986, 16–23
- Gansner, E.R./North, S.C.:** An Open Graph Visualization System and its Applications to Software Engineering. In: Software-Practice and Experience, 0 1999, 1–5
- Gaus, Wilhelm:** Dokumentations- und Ordnungslehre. 2. Auflage. Berlin: Springer, 1995
- Genesereth, M./Fikes, R.:** Knowledge Interchange Format Reference Manual - Version 3.0. Stanford University, 1992 – CSD Tech-Report Logic 92-1
- Geyer-Schulz, Andreas/Hahsler, Michael:** Evaluation of Recommender Algorithms for an Internet Information Broker based on Simple Association Rules and on the Repeat-Buying Theory. In: *Masand, Brij et al. (Hrsg.): Fourth WebKDD Workshop: Web Mining for Usage Patterns & User Profiles.* Edmonton, Canada, Juli 2002, 100–114
- Giugno, Rosalba/Lukasiewicz, Thomas:** P-SHOQ(D): A Probabilistic Extension of SHOQ(D) for Probabilistic Ontologies in the Semantic Web. In: Proceedings of the 8th European Conference on Logics in Artificial Intelligence (JELIA'02). September 2002., 86–97
- Gruber, Thomas R.:** A Translation Approach to Portable Ontology Specifications. In: Knowledge Acquisition, 5 1993, Nr. 2, 199–220, ISSN 1042–8143
- Guarino, N.:** Formal Ontology and Information Systems. In: *Guarino, N. (Hrsg.): Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98.* Trento, Italy: IOS Press, June 1998, 3–15
- Guarino, Nicola:** Formal Ontology, Conceptual Analysis and Knowledge Representation. In: International Journal of Human-Computer Studies, 43 1995, Nr. 5-6, 625–640, ISSN 1071–5819
- Gurzki, Thorsten/Hinderer, Henning/Eberhard, Claus-T.; Bullinger, Hans-Jörg (Hrsg.):** Marktübersicht Portal Software - Marktübersicht für Business-, Enterprise-Portale und E-Collaboration. Stuttgart: Fraunhofer IRB, Februar 2002 (URL: http://www.media-vision.iao.fhg.de/downloads/Portal_Software.pdf)
- Haarslev, Volker/Möller, Ralf:** RACER System Description. In: Proceedings of the First International Joint Conference on Automated Reasoning. Springer-Verlag, 2001, ISBN 3–540–42254–4, 701–706
- Hacker, Rupert:** Bibliothekarisches Grundwissen. München: K.G. Saur, 1992
- Harms, Ilse/Luckhardt, Heinz-Dirk:** Virtuelles Handbuch Informationswissenschaft. 2001 (URL: <http://is.uni-sb.de/studium/handbuch/index.php>)
- Hausser, Roland:** Grundlagen der Computerlinguistik: Mensch-Maschine-Kommunikation in natürlicher Sprache. Berlin, Heidelberg, New York: Springer-Verlag, 2000
- Hearst, Marti:** SIGIR: Call For Participation. University of California, Berkeley, 1999 (URL: <http://www.sims.berkeley.edu/research/conferences/sigir99/old/themes.html>)
- Heflin, J./Hendler, J.:** Searching the Web with SHOE. In: Artificial Intelligence for Web Search. Papers from the AAAI Workshop. WS-00-01. AAAI Press, 2000, 35–40
- Helsper, E.M./van der Gaag, L.C.:** Building Bayesian Networks through Ontologies. In: *Harmelen, F. van (Hrsg.): Proceedings of the 15th European Conference on Artificial Intelligence.* Amsterdam, Niederlande: IOS Press, 2002, 680–684

- Hendler, J.:** Agents and the Semantic Web. In: IEEE Intelligent Systems Journal, 16 März/April 2001, Nr. 2, 30–37
- Hendler, J./McGuinness, D. L.:** The DARPA Agent Markup Language. In: IEEE Intelligent Systems Journal, 15 November/Dezember 2000, Nr. 6, 67–73
- Heyer, Gerhard et al.:** Learning Relations using Collocations. In: *Maedche, A. et al. (Hrsg.):* Proceedings of IJCAI Workshop on Ontology Learning. Seattle/ WA, USA, August 2001
- Horrocks, I.:** The FaCT System. In: *Springer (Hrsg.):* Proceedings of tableaux'98, International Conference on Automated Reasoning with Analytic Tableaux and Related Methods. Saratoga Springs, 1998, ISBN 3–540–66086–0, 307–312
- IEEE Computer Society:** IEEE 1074-1995: Standard for Developing Software Lifecycle Processes. 1995
- International Organization for Standardization:** ISO 9241: Ergonomic requirements for office work with visual display. 1998
- International Organization for Standardization:** ISO 13407: Human-centred Design Processes for Interactive Systems. Genf, Schweiz, 1999
- International Organization for Standardization:** ISO 13250: Information technology. SGML applications. Topic maps. 2000
- International Organization for Standardization:** ISO 15836: Information and documentation - The Dublin Core metadata element set. 2003
- Jackson, Peter/Moulinier, Isabelle:** Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. John Benjamins Publishing Co., 2002
- Jones, D./Bench-Capon, T./Visser, P.:** Methodologies for Ontology Development. In: Proceedings of IT-KNOWS Conference, XV IFIP World Computer Congress. Budapest, August 1998
- Jones, Steve:** Graphical Query Specification and dynamic Result Previews for a Digital Library. In: Proceedings of the 11th annual ACM symposium on User interface software and technology. ACM Press, 1998, ISBN 1–58113–034–1, 143–151
- Jones, William P./Furnas, George W.:** Pictures of Relevance: A Geometric Analysis of Similarity Measures. In: Journal of the American Society for Information Science, 38 1987, Nr. 6, 420–442
- Kaiser, Alexander:** Computer-unterstütztes Indexieren in Intelligenten Information Retrieval Systemen. Ein Relevanz-Feedback orientierter Ansatz zur Informationserschließung in unformatierten Datenbanken. Dissertation, Wirtschaftsuniversität Wien, Wien, 1993
- Karp, P.D./Chaudhri, V.K./Thomere, J.:** XOL: An XML-Based Ontology Exchange Language. Pangaea Systems and SRI, International, 1999 – Technischer Bericht (URL: <http://www.ai.sri.com/~pkarp/xol/>)
- Kifer, Michael/Lausen, Georg/Wu, James:** Logical Foundations of Object-Oriented and Frame-Based Languages. In: Journal of the ACM, 42 1995, Nr. 4, 741–843, ISSN 0004–5411
- Koller, Daphne/Levy, Alon Y./Pfeffer, Avi:** P-CLASSIC: A Tractable Probabilistic Description Logic. In: Proceedings of the AAAI Fourteenth National Conference on Artificial Intelligence. 1997, 390–397
- Koller, Daphne/Pfeffer, Avi:** Object-Oriented Bayesian Networks. In: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97). 1997, 302–313

- Koller, Daphne/Pfeffer, Avi:** Probabilistic Frame-Based Systems. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence. 1998, 580–587
- Kopperger, Dietmar/Schulte, Tamara:** Knowledge meets Process - Wissen und Prozesse managen im Intranet. Stuttgart: Fraunhofer IRB, 2001.– Kapitel Anforderung an Geschäftsprozessmanagement-Werkzeuge, 33–38
- Kuhlen, Rainer:** Zum Stand pragmatischer Forschung in der Informationswissenschaft. In: Pragmatische Aspekte beim Entwurf und Betrieb von Informationssystemen. Proceedings des 1. Internationalen Symposiums für Informationswissenschaft. Universitätsverlag Konstanz, 1990
- Lagus, Krista et al.:** WebSOM for Textual Data Mining. In: Artificial Intelligence Review, 13 1999, Nr. 5-6, 345–364
- Lamping, J./Rao, R.:** Laying Out and Visualizing Large Trees Using a Hyperbolic Space. In: Proceedings of the ACM Symposium on User Interface Software and Technology. ACM Press, 1994, 13–14
- Lancaster, F.W.:** Information Retrieval Systems: Characteristics, Testing and Evaluation. In: Wiley, New York 1968
- Langenscheidt:** Fremdwörterbuch online. Langenscheidt Fachverlag, 2004 (URL: <http://www.langenscheidt.de/fremdwb/>)
- Le Grand, B./Soto, M.:** Information management - Topic maps visualization. In: Proceedings of XML Europe 2000. 2000
- Lenat, D./Guah, R.:** Building large Knowledge-based Systems. Representation and Inference in the Cyc Project. Menlo Park: Addison-Wesley, 1990
- Lewis, David D.:** Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In: ECML '98: Proceedings of the 10th European Conference on Machine Learning. Springer-Verlag, 1998, ISBN 3–540–64417–2, 4–15
- Liddy, Elizabeth D.:** Enhanced Text Retrieval Using Natural Language Processing. In: ASIS Bulletin, 1998 (URL: <http://www.asis.org/Bulletin/Apr-98/liddy.html>)
- Lloyd, J.W./Topor, R.W.:** Making Prolog more expressive. In: Journal of Logic Programming, 3 1984, Nr. 1, 225–240
- Luhmann, Niklas:** Soziale Systeme. Grundriss einer allgemeinen Theorie. Frankfurt a.M.: Suhrkamp, 1984
- Luhmann, Niklas:** Die Wissenschaft der Gesellschaft. Frankfurt a.M.: Suhrkamp, 1990
- Luhn, H.P.:** A statistical Approach to mechanized Encoding and Searching of Library Automation. In: IBM Journal of Research and Development, 1957, 309–317
- Lukasiewicz, Thomas:** Probabilistic Logic Programming with Conditional Constraints. In: ACM Transactions on Computational Logic (TOCL), 2 2001, Nr. 3, 289–339, ISSN 1529–3785
- Luke, S./Heflin, J.:** SHOE 1.01. Proposed Specification. SHOE Project, Februar 2000 – Technischer Bericht (URL: <http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.htm>)
- MacGregor, Robert M.:** Inside the LOOM Description Classifier. In: SIGART Bulletin, 2 1991, Nr. 3, 88–92, ISSN 0163–5719
- Maedche, Alexander/Staab, Steffen:** Mining Ontologies from Text. In: *R.Dieng/Corby, O (Hrsg.):* Proceedings of 12th International Conference on Knowledge Engineering and Knowledge Management, EKAW-2000. Juan-les-Pins, France: Springer, Oktober 2000

- Mannecke, Hans-Jürgen:** Klassifikation. In: Grundlagen der praktischen Information und Dokumentation. München: K. G. Saur, 1997, 141–159
- McBride, Brian:** Jena: A Semantic Web Toolkit. In: IEEE Internet Computing, 6 2002, Nr. 6, 55–59, ISSN 1089–7801
- McCrickard, S.D./Kehoe, C.M.:** Visualizing Search Results using SQWID. In: Proceedings of the Sixth International World Wide Web Conference. April 1997 (URL: <http://www.cc.gatech.edu/grads/m/Scott.McCrickard/sqwid/Doc/www6.html>)
- McGuinness, Deborah L./Harmelen, Frank van:** OWL Web Ontology Language Overview. W3C, März 2003 – Working Draft 31 (URL: <http://www.w3.org/TR/owl-features/>)
- Mikheev, A./Finch, S.:** A Workbench for Finding Structure in Texts. In: Proceedings of the Applied Natural Language Processing, ANLP-97. Washington D.C., USA, April 1997
- Miller, G.A. et al.:** WordNet: An On-Line Lexical Database. In: International Journal of Lexicography, 3 1990, Nr. 4, 235–312
- Minsky, M.:** Semantic Information Processing. Cambridge, USA: MIT Press, 1968
- Minsky, M.:** A Framework for Representing Knowledge. In: *Winston, P.H. (Hrsg.): The Psychology of Computer Vision.* New-York: McGraw-Hill, 1975, 211–277
- Motik, Boris/Maedche, Alexander/Volz, Raphael:** A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications. In: On the Move to Meaningful Internet Systems - Confederated International Conferences DOA, CoopIS and ODBASE 2002. Springer-Verlag, 2002, ISBN 3–540–00106–9, 1082–1099
- Nardi, D./Brachman, R. J.:** An Introduction to Description Logics. In: *Baader, F. et al. (Hrsg.): The Description Logic Handbook.* Cambridge University Press, 2002, 5–44
- Neches, R. et al.:** Enabling Technology for Knowledge Sharing. In: AI Magazine, 3 1991, Nr. 12, 36–56
- Nie, Jian-Yun:** Towards a Probabilistic Modal Logic for Semantic-Based Information Retrieval. In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, 1992, ISBN 0–89791–523–2, 140–151
- Noller, Stephan:** Mentale Modelle und Webnavigation: Ein Usability-Experiment zur Informationssuche im World-Wide-Web. Dissertation, Universität zu Köln, Fachbereich Psychologie, Köln, 2000
- North, Kurt:** Wissensorientierte Unternehmensführung: Wertschöpfung durch Wissen. Wiesbaden: Gabler Verlag, 1998
- Nottelmann, Henrik/Fuhr, Norbert:** pDAML+OIL: A probabilistic extension to DAML+OIL based on probabilistic Datalog. In: Proceedings Information Processing and Management of Uncertainty in Knowledge-Based Systems. 2004
- Novak, Jasminko/Kunz, Christoph/Wurst, Michael:** Entdeckung und Nutzbarmachung von stillem Wissen in heterogenen Expertengemeinschaften. In: *i-com Zeitschrift für interaktive und kooperative Medien*, 2003, Nr. 3, 18–26, ISSN 1618–162X
- Noy, N. Fridman/Hafner, C.D.:** The State of the Art in Ontology Design: A Survey and Comparative Review. In: AI Magazine, 1997, 53–74
- Noy, N.F. et al.:** Creating Semantic Web Contents with Protege-2000. In: IEEE Intelligent Systems Journal, 16 2001, Nr. 2, 60–71

- Object Management Group:** The Unified Modeling Language (UML) Specification. 2001 (URL: <http://www.omg.org/technology/documents/formal/uml.htm>)
- Oppenheim, A. Leo:** Ancient Mesopotamia: Portrait of a Dead Civilization. Chicago, USA: The University of Chicago Press, 1964
- Page, Lawrence et al.:** The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project, 1998 – Technischer Bericht
- Panyr, Jiri:** Information Retrieval Systeme: State of the Art. In: Handbuch der modernen Datenverarbeitung. 1987, 15–36
- Patience, N./Chalmers, R.:** Unstructured Data Management: the Elephant in the Corner. the451, 2002 – Technischer Bericht (URL: <http://ww2.the451.com/reports/udm.php>)
- Pearl, Judea:** Probabilistic Reasoning in Intelligent Systems: Networks of plausible Inference. Morgan Kaufmann Publishers Inc., 1988, ISBN 0–934613–73–7
- Picot, Arnold/Reichwald, Ralf/Wigand, Rolf:** Die Grenzenlose Unternehmung. Dritte Auflage. Gabler Verlag, 2003
- Pirolli, P./Card, S./Wege, M.:** The Effect of Information Scent on Searching Information Visualizations of large Tree Structures. In: Proceedings of AVI 2000. Palermo, Italien: ACM Press, 2000
- Poetzsch, Elenore:** Information Retrieval: Einführung in Grundlagen und Methoden. Potsdam: Verlag für Berlin-Brandenburg, 1998
- Popper, K. R.:** Die Logik der Sozialwissenschaften. In: *Adorno, T. W. et al. (Hrsg.): Der Positivismusstreit in der deutschen Soziologie.* Berlin: Luchterhand, 1969
- Porr, Bernd:** Systemtheorie und Naturwissenschaft. Eine interdisziplinäre Analyse von Niklas Luhmanns Werk. Deutscher Universitäts-Verlag, 2002
- Porter, M.:** An Algorithm for Suffix Stripping. In: Automated Library and Information Systems, 14 1980, Nr. 3, 130–137
- Probst, Gilbert/Raub, Steffen/Romhardt, Kai:** Wissen managen: wie Unternehmen ihre wertvollste Ressource optimal nutzen. Wiesbaden: Gabler-Verlag, 1997
- Puppe, Frank:** Einführung in Expertensysteme. Berlin Heidelberg: Springer Verlag, 1991
- Quillian, M.:** Semantic Memory. In: *Minsky, M. (Hrsg.): Semantic Information processing.* Cambridge, MA: MIT-Press, 1968, 227–270
- Robertson, G./Mackinlay, J./Card., S.:** Cone trees: Animated 3D Visualizations of hierarchical Information. In: Proceedings of ACM CHI'91. ACM Press, 1991, 189–194
- Robertson, S.E./Spark-Jones, K.:** Relevance Weighting of Search Terms. In: Journal of the American Society for Information Science, 27 1976, 129–146
- Rocchio, J.:** Document Retrieval Systems - Optimization and Evaluation. Dissertation, Harvard Computational Laboratory, Cambridge, 1966
- Roget, P.M.:** Thesaurus of English Words and Phrases. Harmondsworth: Penguin, 1852
- Rosenfeld, Louis/Morville, Peter:** Information Architecture for the World Wide Web: Designing Large-Scale Web Sites. 2. Auflage. O'Reilly, August 2002, ISBN 0–596–00035–9
- Salton, G./Yang, C.S.:** On the Specification of Term Values in automatic Indexing. In: Journal of Documentation, 29 1973, 351–372

- Salton, Gerard:** Automatic Text Analysis. In: *Science* 168 1970, Nr. 335-343
- Salton, Gerard:** The Smart Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall, Inc., Englewood Cliffs, 1971 – Technischer Bericht
- Salton, Gerard/McGill, Michael J.:** Introduction to Modern Information Retrieval. New York, USA: McGraw-Hill, 1983
- Salton, Gerard/McGill, Michael J.:** Information Retrieval - Grundlegendes für Informationswissenschaftler. Hamburg etc.: McGraw-Hill, 1987
- Sarkar, M./Brown, M.H.:** Graphical Fisheye Views of Graphs. In: Proceedings of ACM CHI'92. ACM ACM Press, 1992, 83 – 91.
- Schmaltz, R./Hagenhoff, S.:** Semantic Web Technologien für das Wissensmanagement. Universität Göttingen, Abt. Wirtschaftsinformatik II, 2004 (1). – Arbeitspapier
- Schreyögg, Georg/Geiger, Daniel:** Kann implizites Wissen Wissen sein? In: *Bresser, Rudi/Krell, Gertraude/Schreyögg, Georg (Hrsg.):* Diskussionsbeiträge des Instituts für Management des Instituts für Management. Band 14, 2002
- Sebastiani, Fabrizio:** A Note on Logic and Information Retrieval. In: *Ruthven, Ian (Hrsg.):* Proceedings of MIRO-95, Workshop on Multimedia Information Retrieval. Glasgow, UK, 1996, 1–15
- Sebrechts, Marc M. et al.:** Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. In: *Research and Development in Information Retrieval.* 1999, 3–10
- Shannon, C.E./Weaver, W.:** The Mathematical Theory of Communication. Urbana, Illinois: University of Illinois Press, 1949, 379–423 and 623–656
- Soukup, C.:** Wissensmanagement: Wissen zwischen Steuerung und Selbstorganisation. Klagenfurt: Gabler Verlag, 2001
- Sowa, J.:** Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley, 1983, System Programming Series
- Sparck-Jones, Karen:** Automatic Keyword Classification for Information Retrieval. London: Butterworths, 1971
- Sparck-Jones, Karen:** A statistical Interpretation of Term Specificity and its Application in Retrieval. In: *Journal of Documentation*, 28 1972, 111–121
- Sparck-Jones, Karen:** Reflections on TREC. In: Proceedings of the second conference on Text retrieval conference. Pergamon Press, Inc., 1995, 291–314
- Staab, Steffen/Maedche, Alexander:** Ontology Engineering beyond the Modeling of Concepts and Relations. In: Proceedings of the ECAI'2000 Workshop on Applications of Ontologies and Problem-Solving Methods. Berlin, Germany, 2000
- Staud, Josef; Buder, Marianne et al. (Hrsg.):** Wirtschaftsinformation. München, New Providence, London, Paris: Saur Verlag, 1997, 562–565
- Studer, Rudi/Benjamins, V. Richard/Fensel, Dieter:** Knowledge Engineering: Principles and Methods. In: *Data and Knowledge Engineering*, 25 1998, Nr. 1-2, 161–197, ISSN 0169–023X
- Studer, Rudi et al.:** Ontologies and the Configuration of Problem-solving Methods. In: *Gains, B.R./Musen, M.A. (Hrsg.):* Proceedings of the 10th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop. Banff, Canada, 1996, 11–20

- Studer, Rudi/Schnurr, H./Nierlich, A.:** Semantisches Knowledge Retrieval. Ontoprise, Karlsruhe, 2001 – Ontoprise Whitepaper Series
- Trist, E.L./Bamforth, K.:** Some Social and Psychological Consequences of the Longwall Method of Coal Getting. In: *Human Relations*, 4 1951, 3–38
- Trist, E.L. et al.:** Organizational Choice, Capabilities of Groups at the Coal Face under changing Technologies. Tavistock Institute of Human Relation, London, 1963 – Technischer Bericht
- Tufte, Edward R.:** The Visual Display of Quantitative Information. Chesire, USA: Graphics Press, 1983
- Turtle, Howard/Croft, W. Bruce:** Inference Networks for Document Retrieval. In: Proceedings of the 13th Annual International ACM SIGIR. ACM ACM Press, 1990, 1–24
- Uschold, Mike/Grüninger, Michael:** Ontologies: Principles, Methods, and Applications. In: *Knowledge Engineering Review*, 11 1996, Nr. 2, 93–155
- van Rijsbergen, Cornelis J.:** Information Retrieval. 2. Auflage. Dept. of Computer Science, University of Glasgow, 1979
- van Rijsbergen, Cornelis J.:** A new theoretical framework for Information Retrieval. In: SIGIR-86. Pisa, Italy, 1986, 194–200,
- van Rijsbergen, Cornelis J.:** A non-classical Logic for Information Retrieval. In: *The Computer Journal*, 29 1986, Nr. 6, 481–485
- Vapnik, Vladimir N.:** The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., 1995, ISBN 0–387–94559–8
- Vossen, G.:** Datenbankenmodelle, Datenbanksprachen und Datenbankmanagement. München, Wien: Oldenbourg Verlag, 1999
- Weibel, S. et al.:** OCLC/NCSA Metadata Workshop Report. OCLC Online Computer Library Center, Dublin, Ohio, 1995 – Technischer Bericht
- Welty, Christopher A./Ferrucci, David A.:** What's in an Instance? RPI Computer Science, 1994 – Technischer Bericht
- Wenger, Etienne:** Communities of Practice: Learning, Meaning, and Identity. New York: Cambridge University Press, 1998
- Winograd, T.:** Understanding Natural Language. Edinburgh: Edinburgh University Press, 1972
- Wolf, Patricia; Merker, Richard/Groth, Torsten (Hrsg.):** Erfolgsmessung der Einführung von Wissensmanagement. Eine Evaluationsstudie im Projekt 'Knowledge Management' der Mercedes-Benz Pkw-Entwicklung der DaimlerChrysler AG. Münster: Verlagshaus Monsenstein und Vannerdat, 2003, Managementkompetenz
- Wolke, Dieter:** Mathematik für Naturwissenschaftler, Skriptum zur Vorlesung. Mathematisches Institut, Albert-Ludwigs-Universität, Freiburg, 2004 (URL: http://home.mathematik.uni-freiburg.de/wolke/mathe_naturwiss_schuster.pdf)
- Wormser-Hacker, Christa:** Evaluierung im Information Retrieval. Herbstschule Information Retrieval, Schwerte, 1998
- Zadeh, L.A.:** Fuzzy Sets and Systems. In: *Systems Theory*, 1965, 29–37
- Ziegler, Jürgen/Kunz, Christoph/Botsch, Veit:** Matrix Browser: Visualisierung und Exploration vernetzter Informationsräume. In: Konferenzband der Mensch und Computer 02. Hamburg, Germany: Teubner Verlag, 2002

curriculum vitae

persönliche Daten

Name	Christoph Daniel Kunz
Geburtsdatum	04.02.1974
Geburtsort	Friedrichshafen
Familienstand	nicht verheiratet
Anschrift	Senefelderstraße 10 70178 Stuttgart
Kontakt	Festnetz: +49 - (0) 711 - 9641192 Mobil: +49 - (0) 179 / 6954567 e-Mail: christoph@ganzanders.net



beruflicher Werdegang

01.08.01 – heute	Institut für Arbeitswissenschaft und Technologiemanagement, Competence Center Human-Computer Interaction Wissenschaftler. Beratungs-, Forschungs- und Entwicklungstätigkeiten in den Bereichen Interaktive Systeme, Informationsvisualisierung, sowie Informationsmanagement und -retrieval.
01.10.04 – 31.01.05	Fachhochschule Reutlingen, Hochschule für Technik und Wirtschaft Dozentur für die Lehrveranstaltung Medientechnik und Programmierung, Studiengang Wirtschaftsinformatik.
01.04.00 – 31.07.01	I-D Media AG, Division LivingScreen Softwareentwickler. Planung, Entwurf und Implementierung eines Frameworks zum personalisierten Verteilen von dynamischen Inhalten mittels einer Push-Technologie.

Bildung

01.08.01 – 16.12.05	Promotion zum Dr.-Ing., Institut für Arbeitswissenschaft und Technologiemanagement, Fakultät Maschinenbau, Universität Stuttgart. Thema: Ein integrierter Ansatz zur wissensbasierten Informationsrecherche.
15.08.93 – 24.02.00	Studium der Elektro- & Informationstechnik, Universität Stuttgart. Vertiefung der Automatisierungs- und Regelungstechnik. Abschluss als Diplomingenieur. Diplomarbeit: Steuerung und Visualisierung eines industriellen Kaffeeautomaten mit JavaBeans. Studienarbeit: Integration eines Ultraschall-Ortungssystems in die Steuerungssoftware eines mobilen Fahrroboters in C.
1980 – 1993	Schulausbildung zur Erlangung der allgemeinen Hochschulreife am Graf Zeppelin Gymnasium, Friedrichshafen. Leistungsfächer: Mathematik, Physik.