Knowledge Discovery durch Text Mining

Einsatz intelligenter Systeme zur Akquisition, Darstellung und Verteilung von textbasiertem Wissen

Diplomarbeit
Studiengang Informationswirtschaft

Fakultät für Informations- und Kommunikationswissenschaft
Fachhochschule Köln
University of Applied Science Cologne

vorgelegt am 2004-07-05 von Stefan Koch
Matrikelnummer: 11023811
email: diplom@stefkoch.de
www: http://stefkoch.de/diplom/

Prof. Dr.phil. K. Lepsky
Erstprüfer

Prof. Dipl.-Math. W. Gödert
Zweitprüfer

"Like distant islands sundered by the sea,
We had no sense of one community.
We lived and worked apart and rarely knew
That others searched with us for knowledge, too. [...]
But, could these new resources not be shared?
Let links be built; machines and men be paired!
Let distance be no barrier! They set
That goal: design and built the ARPANET! [...]
The second node, the NIC, was soon installed.
The Network Info Center, it was called.
Hosts and users, services were touted:
To the NIC was network knowledge routed."[1]

"The first phase of the Web is
human communication through shared knowledge.
We have a lot of work to do before we have
an intuitive space in which we can
put down our thoughts and build our understanding of
what we want to do and how and why we will do it.
The second side to the Web, yet to emerge,
is that of machine-understandable information.
As this happens, the day-to-day mechanisms of
trade and bureaucracy will be handled by
agents, leaving humans to provide
the inspiration and the intuition."[2]

1 [Cerf1969]
2 [Berners-Lee1997]

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation der Arbeit

Die Nutzung von Informationsressourcen bzw. der tägliche Umgang mit diesen wird in der politischen, administrativen, wirtschaftlichen und wissenschaftlichen Welt heute als Selbstverständlichkeit angesehen, wobei der überwiegende Teil der Informationen in Textform vorliegt.[3] Durch die wachsende Anzahl an Informationsquellen wird das Paradigma des "wachsenden Weltwissens" genährt. Diese Vorstellung basiert auf der Annahme, dass menschliches Wissen in Datenbanken, Dokumenten o.ä. festgehalten werden könne. Zusätzlich ermöglichen Internettechnologien scheinbar die *Ubiquität*[4] von Informationen, eine schnelle Übertragung und somit die unmittelbare und unkomplizierte Nutzung dieses Wissens.[5] Diesen Annahmen entsprechend nehmen die Erwartungen an die Informationssysteme zu. Von sog. *Agentensystemen* und *Management Information Systemen* (MIS) wird die selbstständige Generierung und Aufbereitung von Wissen erwartet. Die Entstehung der sog. *Wissensgesellschaft* scheint dies zur Voraussetzung zu haben. Vor diesem Hintergrund gewinnen Hilfsmittel zum Umgang mit den Wissensquellen zunehmend an Bedeutung.

Sowohl die Entwickler als auch die Nutzer von Informationssystemen werden mit einigen grundlegenden Schwierigkeiten konfrontiert. Zum Einen stellt sie der Umfang der nutzbaren internen sowie der externen Informationsquellen vor ein Auswahlproblem. 1964 bereits als "Megabitbombe" bezeichnet,[6] spricht man heutzutage von dieser Problematik als "Informationsflut" oder "information overload". Da es unmöglich ist, sich einen detaillierten Überblick über alle verfügbaren Informationsquellen zu verschaffen und sie zu vergleichen, muss das Informationssystem die relevanten Informationen **extrahieren** (s. Kapitel 3) und diese im Zusammenhang präsentieren. Uneinheitliche Informations-strukturen[7], Dokumentformate und Datenbankabfragesprachen erschweren diese Aufgabe.

---

3 "Die große Masse an Informationen in unstrukturierten Daten steckt ganz einfach in Texten, i.e. in Textdokumenten, die sich in den individuellen Anwendungen am Arbeitsplatz oder in abteilungs- oder unternehmensweiten Archiven und Dokumenten-Verwaltungslösungen befinden", in: [Martin1998], S.416

4 ubiquitär = überall und zu jeder Zeit verfügbar

5 Diese Zielsetzung wurde bereits für das ARPAnet gesetzt

6 [Lem1964]

7 In vielen Firmennetzen wird kein einheitliches System zur Ordnung und zum Zugriff auf Datenbanken oder Dokumente verwendet.

Eine weitere Problematik stellt die (sowohl in Dokumenten, als auch in Suchanfragen verwendete) natürliche Sprache dar. Beim Auffinden relevanten Wissens müssen diese Formulierungen **analysiert** (s. Kapitel 4) und **abgebildet** (s. Kapitel 5) werden. Ohne geeignete Hilfsmittel z.B. bei der Indexierung nehmen mit wachsender Datenmenge zwar auch die relevanten Daten zu, diese werden aber immer schwieriger auffindbar. Für eine nachhaltige Nutzung der Wissensquellen müssen **standardisierte** Verfahren (s. Kapitel 6) verwendet werden, damit der **Transfer** von Wissen (s. Kapitel 7) sowie dessen **Integration** in die eigene Informationsinfrastruktur (s. Kapitel 8) möglich wird.

1.2 Zielsetzung

Bezüglich o.g. Aufgabenfeldern soll in dieser Arbeit aufgezeigt werden, welchen Beitrag die Verfahren des *Text Mining* (**TM**) bei der Analyse und Strukturierung von Wissen leisten können. Dies soll Entwicklern von *Knowledge Discovery* (**KD**)-Systemen als Entscheidungshilfe dienen und Anwendern ermöglichen, entsprechende Anwendungen auswählen zu können. Hierfür werden die Unterschiede der Verfahren in Bezug auf ihre Ausdruckskraft aufgezeigt. Die Vorstellung der Verfahren soll dabei der Beantwortung folgender Fragestellungen dienen:

- Wie kann aus Textinformationen Wissen *extrahiert* werden?
- Wie kann dieses Wissen *abgebildet* werden?
- Welchen Beitrag kann in diesem Zusammenhang *XML*[8] leisten?
- Welche Verfahren existieren für den *Transfer* von Wissen?
- Wie kann das Wissen nutzbar gemacht bzw. *integriert* werden?

8 XML: Extensible Markup Language, http://www.w3.org/XML/

1.3 Eingrenzung

In der vorliegenden Arbeit werden Wege und Modelle der Analyse von Wissen und Hilfsmittel für dessen Darstellung aufgezeigt. Mit ihrer Hilfe kann die Anzahl der relevanten Dokumente reduziert und somit das *Retrieval* erleichtert werden. Aufgrund der großen Anzahl an Text Mining Anwendungen[9] wird hier der Schwerpunkt auf die theoretischen Verfahren gelegt. Da nicht alle Methoden betrachtet werden können, die der Knowledge Discovery und dem Text Mining zugerechnet werden,[10] richtet sich die Untersuchung hauptsächlich auf die grundlegenden Ansätze. Näher eingegangen wird auf die XML-basierte Textanalyse und auf *semantische Netze*, die den Hintergrund für ein Retrievalsystem und für die Herleitung von Wissen darstellen können.

Die Möglichkeiten der automatisierten Nutzung und Integration der Text Mining-Konzepte in Betriebsprozesse stellen ein weites Feld dar und verlangen jeweils individuell angepasste Umsetzungen. Daher kann dieses Thema hier nicht umfassend bearbeitet werden. Es werden allerdings ein paar Hinweise für die Integration von *Ontologien* in eine Knowledge Discovery Umgebung gegeben. Bei der Betrachtung der gewählten Ansätze, Verfahren und Standards wird der Schwerpunkt auf die Verwendungsmöglichkeit für das Text Mining gelegt.

Folgende Fragestellungen werden bewusst nicht behandelt, da sie zwar im Umfeld von KD und TM eine Rolle spielen, aber keine Kernproblematiken dieser Bereiche darstellen:

- Vorauswahl:
  Welche Quellenauswahl soll im Vorhinein getroffen werden?
- Kommunikationsprobleme:
  Wie kann implizites, menschliches Wissen abgebildet (externalisiert) werden?
- Politisches Problem:
  Wie kann die Nutzung von wissensbasierten Systemen gefördert werden?
- Darstellungsproblem:
  Wie können die Informationen visuell aufbereitet werden?

9 [Gemert2000] stellt 71 Text Mining Tools vor und vergleicht sie.

10 Zu den zum Text Mining beitragenden Wissenschaften zählen u.a. die Informationswissenschaften, Informatik, Statistik und Linguistik.

1.4 Vorgehensweise

Die Bearbeitung des eingegrenzten Themenfeldes der Arbeit wird in folgende Teilschritte gegliedert:

Zunächst werden einige im Text Mining häufig verwendete Begriffe erläutert, um eine Basis für die folgenden Ausführungen zu schaffen.
- Kapitel 2

Anschließend werden dem Text Mining zugerechnete Ansätze der Extraktion von Wissen aus Texten betrachtet.
- Kapitel 3

Auf die zur Extraktion notwendigen Analyseverfahren wird in Kapitel 4 genauer eingegangen. Es werden Verfahren und Aspekte der Textanalyse betrachtet, die eine Vorbedingung der Darstellung von Wissen sind.
- Kapitel 4

Im nächsten Kapitel werden mögliche Darstellungsverfahren betrachtet, die der Einordnung und Strukturierung von Wissen dienen sollen. Sie stellen mögliche Hilfsmittel bei der Abbildung von Wissen beim Indexieren sowie beim Retrieval dar.
- Kapitel 5

In Bezug auf die Anwendung dieser Verfahren werden dann einige auf XML aufbauende Standards und Verfahren besprochen, mit deren Hilfe sowohl die Extraktion als auch die Darstellung von Wissen bewerkstelligt werden kann.
- Kapitel 6

Um einen Ansatz von Wissenstransfer auf Basis dieser Webtechnologien zu zeigen, wird anschliessend auf das sog. Semantic Web eingegangen.
- Kapitel 7

Abschließend werden einige Hinweise bezüglich der Implementierung von Text Mining-Anwendungen gegeben.
- Kapitel 8

2 Begriffsklärung

Um eine begriffliche Grundlage für die anschliessenden Betrachtungen zu schaffen, werden im Folgenden einige Begriffe erläutert, die im Kontext von Knowledge Discovery und Text Mining Verwendung finden.

Daten - Informationen - Wissen

Während die Begriffe *Daten*, *Informationen* und *Wissen* umgangssprachlich oft vermischt werden[11], existieren in der Fachliteratur für alle drei Begriffe unterschiedliche und teilweise widersprüchliche Definitionen.[12] Zur Erklärung der verschiedenen Begriffe werden verschiedene Abgrenzungsmerkmale herangezogen.[13] Die Begriffe Daten und Informationen lassen sich durch Darstellung des Prozesses der Informationsübertragung voneinander unterscheiden. Laut H. Wedekind[14] besteht dieser aus folgenden Phasen:

Physikalische Signale
↓Ordnungskriterium (Syntax)
Daten
↓Bedeutung (Semantik)
Nachrichten
↓Zweckorientierung (Pragmatik)
Informationen

Informationen sind demnach bereits bezüglich Bedeutungsinhalt und Verwendungskontext interpretiert. Neben der Einordnung und Zuweisung von Bedeutung wird dieser Prozess auch als Filterprozess verstanden, bei dem Informationen aus Daten gewonnen werden. Wird dieser Prozess fortgeführt, so entsteht aus den Informationen *Wissen*. Dieses Modell geht also davon aus, dass beim Entstehungsprozess von Wissen keine Daten oder Informationen hinzukommen, sondern diese lediglich gefiltert werden und Wissen letztendlich eine "übergebliebene" Datenmenge darstellt. T. Davenport geht davon aus, dass beim Entstehungsprozess von Wissen neben der Filterung der Informationen ein Abgleich mit gespeicherten Informationen (Erfahrungen) notwendig ist.

11 Vgl. http://www.net-lexikon.de/Wissen-Begriffsklaerung.html: Wissen = Information.
12 [Stenmark2001], [Aamodt1995]
13 Flüchtigkeit (Daten sind am flüchtigsten, Wissen am beständigsten)
Interpretation/Kontext (Wissen ist kontextbezogen, Daten müssen erst interpretiert werden).
14 [Wedekind1998]

Im Rahmen der sog. *Kontextanalyse* ergebe sich ein Netz von Beziehungen, das als Wissen bezeichnet wird:

> "Wissen ist eine fließende Mischung aus strukturierten Erfahrungen, Wertvorstellungen, Kontextinformationen und Fachkenntnissen, die in ihrer Gesamtheit einen Strukturrahmen zur Beurteilung und Eingliederung neuer Erfahrungen und Informationen bietet. Entstehung und Anwendung von Wissen vollzieht sich in den Köpfen der Wissensträger."[15]

Bezüglich des Aspektes der Veränderung ("fließende Mischung") hebt W. Stock hervor, dass Wissen im Vergleich zu Informationen einen statischen Charakter besitzt.[16]

Während C. Shannon Wissen per Definition als Reduzierung von Unsicherheit bezeichnet,[17] existiert im Widerspruch hierzu auch der Begriff *unsicheres Wissen*, mit dem ungewisses, unzuverlässiges, ungenaues, unscharfes oder unvollständiges Wissen gemeint ist. Dementsprechend wird es durch Wahrscheinlichkeits-, Näherungs-, Ersatz-, Hilfs- oder Durchschnittswerte dargestellt.

Menschliches Wissen wird oft auch als *implizites*, *subsymbolisches* Wissen oder *tacid knowledge* bezeichnet. Man spricht hingegen von *explizitem* oder *externalisiertem* Wissen, wenn das "lebendige" (und deshalb dynamische), stille (kognitive) Wissen in materielle Träger (Artefakte) wie Dokumente, Datenbanken oder Netze verkörpert und damit sozusagen "eingefroren" wird.[18] Neben den hier gezeigten Interpretationsmöglichkeiten existiert für den Wissensbegriff eine Vielzahl anderer Definitionen aus verschiedenen Wissenschaftsbereichen, die hier nicht weiter diskutiert werden können. Stattdessen werden aus den obigen Wissensmodellen wesentliche Aspekte herausgezogen, die im Weiteren als hinreichend für die Existenz, Erstehung (Emergenz) bzw. Identifikation von Wissen und dessen Darstellung angesehen werden: Ein System, welches Informationen extrahiert (filtert), strukturiert und in einen Kontext einbettet und damit der Beurteilung von Informationen dient, kann als Wissen angesehen werden.

15 [Davenport1998]
16 [Stock2000a], S.41
17 [Shannon1980]
18 Es wird dann immer noch Wissen genannt. Auch Methoden, Vorgehensweisen und Strukturen können abgebildet werden, und zwar in programmierter Form.

Knowledgebase

Allgemein kann eine *Knowledgebase* als ein System verstanden werden, in dem Wissensobjekte hinterlegt und gesucht werden können. Neben einer spezifischen mathematischen Definition[19] existieren für den Begriff verschiedene Sichtweisen und Anwendungsfelder. Die "Microsoft knowledge base"[20] z.B. ist eine sequentielle Ansammlung[21] von FAQs bzw. Hilfsdokumenten und erfüllt zwar o.g. Anforderungspunkte an Wissen, die mathematische Definition aber nicht (u.a. fehlende Ontologie). Es gibt allerdings auch Knowledgebases, deren Spektrum an Funktionalitäten über die Darstellung von Texten hinausgehen und verschiedene Struktur- und Abfragemechanismen aufweisen, die die Nutzung von Wissen unterstützen. Einige solcher Plattformen werden in Kapitel 3 vorgestellt.

Knowledge Discovery (KD)

Knowledge Discovery umfasst das Auffinden (Orten), Strukturieren und Dokumentieren von neuem oder vorhandenem Wissen in Datensammlungen oder einer Knowledgebase. Neues Wissen wird dabei mit Inferenzregeln hergeleitet und vorhandenes Wissen mithilfe Extraktion geortet. Für Beides sind Analyse und Darstellung des Wissens Voraussetzung. Aus den extrahierten Strukturen soll menschen-verständliches Wissen generiert werden, das direkt weiterverwendbar sein soll. Hervorgegangen ist der Begriff aus dem Umfeld des *Data Mining*. Während im Data Mining-Umfeld von *Knowledge Discovery in Databases* (**KDD**)[22] gesprochen wird, bezeichnet man den Vorgang im Text Mining als *Knowledge Discovery in Text* (**KDT**)[23].

19 Mathematische Definition: "A knowledge base is a structure KB := (O, I, Vc, Vr) consisting of
an ontology O := (C, <, R, o, A),
a set I whose elements are called instance identifiers (or instances or objects for short),
a function Vc:C -> P(I) called concept instantiation,
a function Vr:R -> P(I x I) called relation instantiation."
aus: [Sure1999], S.48-49

20 http://support.microsoft.com/

21 Es werden fortlaufende Nummern der Artikel vergeben

22 [Fayyad1996a]

23 [Rajman1997], S.4

Data Mining

Data Mining wird meist als Oberbegriff für mehrere Techniken verwendet. Nach G. Nakhaeizadeh[24] ist Data Mining im Prinzip eine Erweiterung des von Codd begründeten On-Line Analytical Processing (OLAP)[25]. Bereits OLAP verhelfe dem Menschen dazu, Wissen zu generieren, und zwar durch Analyse von Trends in Datenbanken[26] und Darbietung verschiedener Aggregationsformen bzw. Sichtweisen auf die Datenmenge. Das Durchschreiten der Stufen der Abstraktion oder Aggregation wird als "rollup" bzw. "rolldown" oder "drilldown"[27] bezeichnet. Data Mining gehe allerdings noch einen Schritt weiter als OLAP, indem die Wissensgenerierung automatisiert werde. Die Data Mining-Verfahren würden:

- selbständig Hypothesen über Zusammenhänge, Muster und Trends generieren,
- die Hypothesen anhand von Daten überprüfen und
- die gültigen Hypothesen herausfiltern können.

R. Agrawal vom IBM Almaden Research Center[28] setzt, genau wie K. Wilde[29], Data Mining und KDD gleich:

> "Data mining (Stonebraker et al. 1993) (also called knowledge discovery in databases (PiatetskyShapiro & Frawley 1991)) is the efficient discovery of previously unknown patterns in large databases, and is emerging as a major application area for databases (Gartner Group 1994) (Business Week 1994)."[30]

A. Kurz[31] und U. Fayyad/G. Shapiro[32] definieren Data Mining hingegen als Schritt bzw. Mittel, das Ziel der KDD zu erreichen. P. Kischka grenzt Data Mining zu KDD wie folgt ab:

> "Data Mining beschreibt die Anwendung von Algorithmen, um Strukturen (patterns) aus Daten abzuleiten, KDD hat zum Ziel, aus den Daten nützliches Wissen abzuleiten."[33]

---

24 [Nakhaeizadeh1998], S.44f.
25 [Codd1993]
26 [Hönig1998], S.171
27 ebd., S.172
28 http://www.almaden.ibm.com/
29 [Wilde2001]
30 [Agrawal1995]
31 [Kurz1998]
32 "The distinction between the KDD process and the data-mining step (within the process) is a central point of this article." aus: [Fayyad1996b]
33 [Nakhaeizadeh1998], S.152

Text Mining (TM)

Der Begriff *Text Mining* erfüllt zwei Aufgaben. Er soll sowohl eine Anlehnung an den Begriff Data Mining darstellen, als auch gerade den Unterschied zu diesem verdeutlichen. Data Mining bezieht sich in der Regel auf die Untersuchung von Faktendatenbanken, die mit numerischen Einträgen (z.B. Umsatzdaten) gefüllt sind. Gegenstand vom Text Mining ist hingegen die Untersuchung von Textdatenbanken, Textsammlungen bzw. Dokumenten, wobei die Methoden des Data Mining angewendet werden. Dementsprechend wird TM auch als Document Mining bezeichnet.[34]

Die "Daimler Chrysler Research and Technology"-Forschungsgruppe definiert Text Mining als Zweig bzw. Untergruppe von Data Mining mit der Analyse von Text als dessen Hauptaufgabe.[35] Durch diese soll dem Benutzer sowohl der Überblick über die Textmengen als auch der Zugriff auf diese ermöglicht werden. Ersteres wird durch Organisieren der Texte (clustering, classification) und Letzeres durch Auswahl (choice) bzw. Extraktion (extraction) der Texte ermöglicht. Bei der Bewältigung dieser Aufgaben stellen sich die numerisch-orientierten Methoden des Data Mining als nur begrenzt einsetzbar heraus, da sie die Quantifizierung der Texte in irgendeiner Form voraussetzen. Aus diesem Grund werden die im TM einsetzbaren Data Mining-Methoden um weitere linguistische und statistische Verfahren ergänzt. Dem Rechnung tragend benutzt die Deutsche Bibliothek[36] folgende erweiterte Definition:

> "Das Gebiet des T.M. umfasst vielfältige Methoden zur Extraktion von Informationen aus natürlichsprachlichen Texten. Die Methoden stammen aus den Forschungsgebieten Wissensrepräsentation, Maschinelles Lernen, Computerlinguistik, Informationsextraktion, Information Retrieval, Mustererkennung."[37]

Text Mining und Knowledge Discovery

Während TM ein Gebiet beschreibt, in dem klassische informations-wissenschaftliche Verfahren wie *Klassifizierung* und *Retrieval* zum besseren Auffinden von relevanten Texten, den sog. Nuggets[38], verwendet werden, beschreibt KD das Auffinden und Nutzen von Wissenstrukturen. Die Schnittmenge beider Ansätze besteht im Aufbau oder in der Nutzung von semantischen Strukturen, die im Retrieval eingesetzt werden können.
In Anlehnung an die Definitionen von Data Mining könnte Knowledge Discovery in Texten als Ziel des Text Mining oder als gleichbedeutend mit diesem betrachtet werden.

---

34 [Dixon1997]
35 [Franke2003]
36 http://www.ddb.de
37 Quelle: http://z3950gw.dbf.ddb.de NORMDATEN: Schlagwort (4728093-1)
38 In Analogie zu den Nuggets einer Goldmine

Artificial Intelligence (AI)

Einige der KD- und TM-Verfahren bedienen sich u.a. Erkenntnissen aus dem Forschungsgebiet der *Künstlichen Intelligenz* (KI). Die KI stellt einen Forschungsbereich dar, der Ergebnisse aus der Informatik, Psychologie, Philosophie und Linguistik integriert.[39] Das Ziel der KI ist es, Vorgänge in maschinenverständliche Sprache abzubilden, die intelligentes Handeln implizieren:

> "Künstliche Intelligenz (KI) ist die Untersuchung von Ideen, die es Computern ermöglichen, intelligent zu sein."[40]

Die Schwierigkeit, KI zu definieren, liege nach P. Winston hauptsächlich bei dem Begriff Intelligenz. KI sei dann nur noch die Abbildung intelligenter Prozesse in Computersprache. Intelligenz ist nach Winston:

> "[...] eine Mischung vieler informationsdarstellender und -verarbeitender Talente."[41]

Um Maschinen intelligentes Handeln zu ermöglichen, bedarf es neben der Abbildung in Maschinensprache der Verknüpfung möglichst vieler dieser Fähigkeiten. Dementsprechend wird die KI in mehrere Teilgebiete aufgeteilt.[42] Zu den Hauptgebieten zählen:

- Bildverstehen,
- (auditives, visuelles und textbasiertes) Sprachverstehen,
- Wissensrepräsentation und
- Logik.

Einige der in späteren Kapiteln vorgestellten TM-Verfahren wurden im Forschungsumfeld der KI theoretisch begründet.

39 [Rich1983]
40 [Winston1987]
41 ebd.
42 [Nilsson1980]

Semantisches Netz

Ein *semantisches Netz* ist ein Netz mit Begriffen als Knoten und Beziehungen bzw. Bindungsstärken als Kanten.[43] Ziel des Netzes ist es, das *Bedeutungsumfeld* von Begriffen darzustellen. Die Semantik, d.h. die Bedeutung eines Begriffes, wird neben dem Versuch, den Begriff zu definieren, vor allem durch seinen *Verwendungskontext* in der Sprache bestimmt:

> "Wer wissen will, was dieses Wort bedeutet, muss zusehen, wie es gebraucht wird – und dies ist der einzige Weg, Aufschluss über seine Bedeutung zu erlangen."[44]

Der semantische Kontext eines Begriffes ist demnach bestimmt durch die Verwendung von anderen Wörtern in seiner Umgebung. Die unmittelbare Umgebung im selben Satz oder Dokument wird auch als lokaler Kontext bezeichnet, während sich der globale Kontext aus der Relation zu anderen Textkorpora aus der Kollektion bzw. durch Heranziehen externen Wissens ergibt. Die in einem semantischen Netz festgehaltene Relation zweier Begriffe kann auf statistischen Häufigkeitswerten basieren oder eine linguistische Beziehungen der Wörter beschreiben.[45] Ein semantisches Netz wird dementsprechend durch Analyse und Verknüpfen von Dokumenten oder Begriffen erreicht. Es kann dazu benutzt werden, einen Überblick über verwandte Begriffe zu geben (Navigation) und stellt somit eine Auswahlhilfe für Begriffe bzw. die damit verknüpften Dokumente dar (Retrieval):

> "Semantische Netze sind eine leistungsstarke Methodik zur Strukturierung von Informationen. Im Sinne eines verknüpften Meta-Daten-Indices finden sie auch in der Verbindung mit dem Dokumentenmanagement zunehmend Anwendung. Dabei werden einzelne Dokumente mit Begriffen im Netz verknüpft, die wiederum selbst als beschreibendes Merkmal mit beliebig vielen Dokumenten verbunden sind. Anders als in hierarchischen Verzeichnisbäumen sind im Semantischen Netz die Zusammenhänge zwischen Begriffen für die Anwender transparent. Aufgrund der Mehrdimensionalität des Netzes können Menschen in beliebigen Richtungen navigieren und dabei je nach konkretem Bedarf und Kontext verschiedene Wege gehen. Daraus resultiert eine bedeutend höhere Effizienz bei der Informationssuche."[46]

In Kapitel 5 werden einige Ausprägungsformen von semantischen Netzen vorgestellt.

---

43 [Stock2000a], S.142
44 [Wittgenstein1949]
45 Im GermaNet, dem deutschen Äquivalent des englischen WordNet (http://www.cogsci.princeton.edu/~wn) gibt es bspw. Antonymie-, Hyponymie-, Hyperonymie-, Holonymie-, Meronymie-, Bestandteils-, Kausal- oder Ableitungsbeziehungen, siehe http://www.sfs.nphil.uni-tuebingen.de/lsd/
46 [Beier2003]

3 Extraktion

Extraktion von Wissen bedeutet Ausfilterung von Sätzen oder Wörtern aus Dokumenten bzw. Dokumentkollektionen. Rekombiniert man diese Objekte, so wird neues Wissen generiert und man spricht nicht mehr von Extraktion von Wissen. Ist die Textmenge beschafft oder lokal geortet, also ausgefiltert worden, so kann sie analysiert und das Wissen extrahiert werden. Diese Vorgänge stellen die Grundlage von Knowledge Discovery dar.

Das Wissen kann u.a. in "Wissen über Texte" und "Wissen aus Texten", sowie in *formales* und *inhaltliches Wissen* unterschieden werden. Das "Wissen über Texte" wird Metawissen genannt. Dieses Metawissen ("background knowledge"[47]) kann einem Dokument bei dessen Erstellung oder nach dessen Analyse zugeordnet werden. Bei der Suche nach einem Dokument kann es dann als Hilfsmittel verwendet werden. Desweiteren kann das Metawissen als Bezugssystem für Textobjekte verstanden werden, an dem die Texte gemessen und in das sie einsortiert werden können. So ist z.B. die Struktur einer Datenbank oder das Dateisystem, in dem Dokumente abgelegt werden, bereits Metawissen über die darin enthaltenen Objekte. Im Gegensatz dazu liefert das "Wissen aus Texten", das sog. immanente Wissen, Aussagen über den Betrachtungsgegenstand des Textes selber. Mithilfe verschiedener Extraktionsmethoden soll es in maschinenlesbare Form gebracht werden, um es weiterverarbeiten zu können. Soll es also für die Text Mining Anwendung zur Verfügung stehen, so muss dessen Struktur *extrahiert*, *analysiert* und *dargestellt* werden.

Im Folgenden wird erläutert, in welchen digitalen Umgebungen und Formaten textuelles Wissen allgemein vorliegen kann. Nicht-digitale Textmengen und die damit verbundenen Problematiken bezüglich Erschliessung, Aufbereitung und Suche in diesen (OCR-Dokumenten)[48] werden hier ausgeklammert. Es werden einige Schwierigkeiten aufgezeigt, die mit der Extraktion von Wissen aus diesen Umgebungen verbunden sind und anschliessend einige grundlegende Ansätze der Wissensextraktion beschrieben.

47 [Rajman1997], S.7
48 [Myka1996b], [Myka1997]

3.1 Wissensumfeld

Textbasiertes Wissen kann in unterschiedlicher Form vorliegen. Das Spektrum reicht von der einfachen Textdatei über Worddokumente, XML-Dateien, PDF-Dateien, Datenbanken, Lernsysteme bis zu Knowledgebases und Expertensystemen[49].

Dokumente werden üblicherweise in einem Dateisystem mit einer hierarchischen Ordnerstruktur abgelegt. Dabei wird häufig ein im Netzwerk freigegebener Ordner als zentrale Dateiablage benutzt. Die manuell gepflegte Namensstruktur der Dateien und Ordner kann je nach Qualität, Konsistenz und Ausprägung in die Wissensstruktur einbezogen werden.[50] Aufbauend auf diesen Strukturen bieten sog. *Dokumentmanagementsysteme* (DMS) eine Verwaltungshilfe durch Indexierung und Suchmöglichkeiten.

Meist unabhängig vom Dateisystem und in Form einer Datenbank speichert ein *Content Management System* (CMS) die Wissensobjekte ab. In diesem System lassen sich die Inhalte direkt bearbeiten, ohne dass sich der Benutzer um den Ort der Ablage kümmern muss. Einige Unternehmen erweitern diese Funktionalität um die eines Portals mit Personalisierungs- und Such-funktionalitäten. Die wohl bekannteste offene Form eines CMS ist das "Wiki-System"[51], mit der Wikipedia[52] als populäres Beispiel. Die Verknüpfung der einzelnen Dokumente findet hier manuell statt. Das sog. "Twiki-System"[53] erweitert die rudimentären Retrievalfunktionen um reguläre Ausdrücke[54] und ermöglicht inhaltlich komplexere (Klassifikations-)Strukturen[55]. Diese sind allerdings in einem proprietären Format und wären nur durch individuelle Anpassungen verwendbar.

49 Expertensysteme werden hier ausgeklammert, da sie aufgrund ihrer speziellen Einsatzbereiche nur schwer zu verallgemeinern sind. Vgl. [Lusti1990], S.249ff.
50 Es ist allerdings davon auszugehen, dass diesbezüglich meist keine umfassenden, verbindlichen und eindeutigen Konventionen bzw. Richtlinien existieren, sodass hier selten Konsistenz erreicht wird.
51 http://wiki.org
52 http://wikipedia.org
53 http://twiki.org
54 http://twiki.org/cgi-bin/view/TWiki/RegularExpression
55 http://twiki.org/cgi-bin/view/TWiki/TWikiForms

Ein Versuch, Wissen in menschenverständlicher Form abzulegen und für Schulungszwecke zu nutzen, stellen die sog. *Computer Based Training* (CBT), *Web Based Training* (WBT) oder *Learning Frameworks* genannten Systeme dar. Sie stellen Lerninhalte meist in gängigen Dateiformaten zur Verfügung (HTML, PDF o.ä.). Um die Verwaltung, Einordnung und Transfer der Wissensobjekte zu ermöglichen, entwarf das "Aviation Industry CBT Committee" (AICC)[56] Richtlinien zur Darstellung von Lerninhalten, den sog. "AICC Guidelines & Recommendations". Diese stellen einen allgemeinen Qualitätsstandard in Bezug auf Entwurf, Verteilung und Bewertung von Lernmaterialien dar.
Die sog. *Educational Modeling Language* (EML)[57] klassifiziert die Texte des Lernsystems nach ihrer Funktion in diesem System. Die Texteinheit wird bspw. als Lerneinheit (text), Aufgabenstellung (task, assignment) oder Test (test) kategorisiert. Die von EML verwendeten Metabeschreibungen beziehen sich also nur auf die formelle Einordnung von Inhalten, nicht aber auf den Inhalt der Einheiten.

Die Wissensobjekte liegen, wenn nicht in o.g. Systemen, und das scheint häufiger der Fall zu sein, meist in Dokumenten vor. Daher werden im Folgenden einige häufig verwendete Dokumentformate näher betrachtet. Man kann zwischen strukturierten und unstrukturierten Dokumentformaten unterscheiden, wobei es hier Abstufungen gibt. Die Struktur kann formeller oder inhaltlicher Natur sein. Eine Aneinanderreihung von Sätzen (Fließtext) kann zwar eine inhaltliche Struktur besitzen, für einen Automaten (Computer) ist sie allerdings bloß eine unstrukturierte Anreihung von Zeichen. Die Überführung inhaltlicher Strukturen in formelle und damit in eine maschinen-verständliche Sprache stellt eine der wesentlichen Aufgaben des TM dar, auf die später näher eingegangen wird.

56 http://www.aicc.org
57 http://eml.ou.nl

HTML-Dokumente[58] sind von der Programmiersprache her zwar formell gesehen strukturiert und bieten auch die Möglichkeit einer inhaltlichen Strukturierung, erfüllen dieses Kriterium aufgrund der uneinheitlichen Verwendung der Tags (HTML-Elemente) meist jedoch nicht. Ein wesentliches Merkmal von HTML ist dessen Möglichkeit, Hyperlinkstrukturen zu erstellen. Diese Strukturen lassen sich u.a. zur Suche nach ähnlichen Dokumenten verwenden.[59]

Ein weiteres Beispiel für ein formell unstrukturiertes Dokument ist Text im *ASCII*-Format[60]. Hier kann der gesamte Text nur als eine bloße Anreihung von Zeichen interpretiert werden.

In Textverarbeitungsprogrammen wie *Microsoft Word* oder *OpenOffice.org* besteht hingegen die Möglichkeit, durch Abschnitte, Überschriften, Fußnoten, Vergabe von Metainformationen u.v.m. das Dokument inhaltlich zu strukturieren. Ein Brief kann so beispielsweise in die inhaltlichen Einheiten "Absender", "Empfänger", "Betreff" und "Textkörper" unterteilt werden.
Da sowohl Microsoft Office als auch OpenOffice.org die Dokumente intern als XML-Dateien abspeichert, sind diese Kontextinformationen der einzelnen Textteile theoretisch nutzbar.[61] Aufgrund der unterschiedlichen Verwendung der jeweiligen Formatierungen wird diese Aufgabe allerdings erschwert.

Die Nutzung der Dokumentstrukturen einer *PDF*-Datei[62] gestaltet sich noch schwieriger, da dieses Format derzeit nur eine Beschreibung des gesamten Dokumentes durch ein Keyword-Feld zulässt.[63]

Auf die vielfältigeren Möglichkeiten, die das *XML*-Format bietet, das Umfeld eines Wissensobjektes zu beschreiben und damit ein Dokument zu strukturieren sowie durchsuchbar zu machen, wird in Kapitel 6 eingegangen.

58 HTML: HyperText Markup Language
59 [Savoy1991], [DeBra1994]
60 ASCII: American Standard Code for Information Interchange
61 Micosoft Office XML Schema: http://www.microsoft.com/office/xml/default.mspx
OpenOffice.org: Das Dokument wird als komprimierte XML-Dokumentkollektion abgespeichert. (u.a. in den Dateien "meta.xml" und "content.xml")
62 PDF: Portable Document Format
63 PDF ist als Dokumentformat eher aufgrund seiner Stärken in der unverfälschten Darstellung am Bildschirm sowie der originalgetreuen Ausgabe am Drucker so bekannt geworden.

Ein Versuch, eine Umgebung zu schaffen, in der Wissen direkt eingegeben und in maschinenverständlicher Sprache abgespeichert werden kann, stellt die sog. "commonsense knowledge base" der Cycorp Inc.[64] dar. Hervorgegangen ist diese Knowledgebase aus dem "Rapid Knowledge Formation"-Projekt (RKF)[65] der DARPA[66], das eine Fortsetzung des 2000 abgeschlossenen "High-Performance Knowledge-Based Technology"-Projektes (HPKB)[67] darstellt. Mithilfe des RKF-Projektes wurde es sogenannten *subject matter experts* ermöglicht, ohne vorheriges Training in "knowledge representation, acquisition or manipulation" Wissen direkt einzugeben und zu ändern. Die "OpenCyc" genannte Plattform ist die frei zugängliche OpenSource-Version dieser Technologien und soll das Abbilden von Aussagen aus dem Allgemeinwissen ermöglichen:

> "OpenCyc is the open source version of the Cyc technology, the world's largest and most complete general knowledge base and commonsense reasoning engine."[68]

Für die Extraktion aus einer Cyc Knowledge Base muss eine eigene Abfragesprache verwendet werden. Ein anderer Versuch, Allgemeinwissen abzubilden, stellt das "Openmind"-Projekt[69] des MIT dar, welches aber lediglich eine Ansammlung von Sätzen darstellt, die weder linguistisch analysiert noch strukturiert bzw. in sinnvolle Zusammenhänge gebracht werden. Eine im Vergleich zur Cyc Knowledgebase weniger auf abstraktes Allgemeinwissen, sondern auf die Darstellung von Beziehungen realer Objekte bzw. Subjekte abzielende Knowledgebase, stellt TAP[70] dar. Sie liefert z.B. die Aussage, dass George Bush Präsident der USA ist.[71]

64 http://www.cyc.com
65 http://reliant.teknowledge.com/RKF/
66 Die Defense Advanced Research Projects Agency (früher ARPA) ist die zentrale Forschungs- und Entwicklungseinrichtung des US-Verteidigungsministeriums.
67 http://reliant.teknowledge.com/HPKB/
68 http://www.opencyc.org
69 http://commonsense.media.mit.edu
70 http://tap.semanticweb.org, http://tap.stanford.edu/tapkb, [Guha2003]
71 http://tap.semanticweb.org/cgi-bin/kb.pl?node=UnitedStatesPresident

3.2 Information Retrieval

Mithilfe von *Information Retrieval* werden Teile aus einer Textmenge extrahiert, die bestimmte Kriterien erfüllen. So lässt sich mit professionellen Retrievaltools u.a. mit Booleschen Operatoren, Gewichtungsfaktoren, Abstandsoperatoren nach ähnlichen Dokumenten suchen. Mithilfe von Retrievalsprachen lassen sich bestimmte Bereiche der Wissensbasis extrahieren und mit anderen Bereichen in Beziehung setzen.

Mit den Retrievalverfahren im internationalen Bereich setzt sich seit 1992 die jährliche von der DARPA ins Leben gerufene "Text Retrieval Conference" (TREC)[72] auseinander. Im deutschen Bereich zählt die "German Indexing and Retrieval Testdatabase" (GIRT) als vergleichendes Instrument für Retrievaltests[73]. Die "Special Interest Group on Information Retrieval" (SIGIR)[74], ins Leben gerufen von der "Association for Computing Machinery" (ACM)[75], richtet ebenfalls eine jährliche Konferenz zum Information Retrieval aus. Mit der Anwendung intelligenter Technologien und dementsprechender Forschung im Bereich Information Retrieval befasst sich u.a. das "Center for Intelligent Information Retrieval" (CIIR)[76] der Universität Massachusetts.

Es stehen mittlerweile etliche Retrievalsysteme zur Verfügung, die für die Suche in Datenbanken oder Dokumenten verwendet werden können. Suchmaschinen für Internet und Intranet beherrschen mittlerweile den Umgang mit verschiedenen Dokumenttypen.[77] Retrieval-Methoden können neben der Suche in Datenbanken oder Textkollektionen auch auf die erwähnten Knowledgebases angewendet werden. Das Retrieval kann hierbei insbesonders durch Verfahren der AI unterstützt werden.[78] Für diesen Zweck wurde 1993 von der DARPA die sog. *Knowledge Query and Manipulation Language* (**KQML**)[79] entwickelt. Sie stellt eine Retrieval-Erweiterung für Knowledgebases dar. Sie wurde allerdings bislang hauptsächlich als Agentenkommunikationssprache eingesetzt.

---

72 http://trec.nist.gov
73 Retrievaltests testen z.b. Recall und Precision-Ergebnisse von diversen Rankingalgorithmen, wobei diese Werte für die Testmenge (bzw. die Treffermenge der vorgegenen Aufgabenstellung) bereits feststehen und intellektuell überprüft wurden.
74 http://www.sigir.org
75 http://www.acm.org
76 http://ciir.cs.umass.edu
77 Die von Google indexierten Dokumenttypen sind auf http://aset.its.psu.edu/googledocs/filetypes.html aufgelistet.
78 [Wong1986]
79 http://www.cs.umbc.edu/kqml/, http://www.csee.umbc.edu/kqml/

3.3 Information Extraction

Während man im informationswissenschaftlichen Bereich vom Information Retrieval spricht und damit die menschliche Tätigkeit meint, Dokumente zu suchen, versteht man unter *Information Extraction* den Vorgang der automatisierten Filterung bzw. Entnahme von Informationen oder "Facts"[80]. Solche Facts sind z.B. Börsendaten, Personenangaben oder Zeit- und Ortsangaben. Ziel dieser Extraktion ist u.a. eine Vereinheitlichung der Daten:

> "A prime motivation is the warehouse address cleaning problem of transforming dirty addresses stored in large corporate databases as a single text field into subfields like "City" and "Street". [...] A second motivating example is cleaning bibliographic records for the construction of citation indices like Citeseer."[81]

Im TM wird Information Extraction u.a. dazu eingesetzt, die Analyse der Texte vorzubereiten:

> "In the proposed framework for text mining, IE plays an important role by preprocessing a corpus of text documents in order to pass extracted items to the data mining module."[82]

Neben Ansätzen, aus unstrukturierten Texten Fakten zu extrahieren, besteht auch die Möglichkeit, bestehende Dokumentstrukturen wie z.B. die Baumstruktur von HTML-[83] und XML-Dateien[84] zu verwenden. Auch die Verwendung verschiedener Tags (z.B. Titel- oder Metatags) sowie Formatangaben zur Schriftgröße o.ä. kann bei der Extraktion berücksichtigt werden.

80 [Maynard2003]
81 [Borkar2001]
82 [Mooney2002]
83 [Freitag1998]
84 [Kosala2002]

Nach der Extraktion von Datums- und Ortsangaben aus einem Text kann eine Zuordnung zu anderen Text-Objekten hergestellt werden, die sich auf denselben Ort oder Zeitraum beziehen. So können chronologische oder geographische Aussagen über die Gesamtheit oder Teile von Texten getroffen werden. Statistische Analysen implizieren z.B. aufgrund häufigen gemeinsamen Auftretens der Worte "Präsident" und "Bush" eine Korrelation der Begriffe für den betreffenden Zeitraum. So könnte aufgrund von Sätzen wie "Bush war in London am ..." und "Am ... war Bush in Kabul" auch aus unterschiedlichen Dokumenten eine Zeitreihe für Bush erstellt werden. Für die Zuverlässigkeit einer solchen Zeitreihe müsste allerdings abgeklärt werden, dass "Bush" auf dieselbe "Instanz Mensch" verweist.[85] Aufgrund von Häufigkeiten kann das System nur Wahrscheinlichkeitswerte in Bezug auf den Wahrheitsgehalt dieser Aussage angeben.[86]

Ein Beispiel für eine Information Extraction-Anwendung ist die OpenSource-Anwendung ANNIE[87]. Sie hebt alle aus vorgegebenen WWW-Seiten extrahierten Textstellen farblich hervor, die den Kategorien "Person", "Location", "Organization", "Date", "Address", "Money" oder "Percent" zugeordnet werden konnten.

Um Information Extraction zu beschleunigen, kann durch Verfahren wie Klassifikation (s. Kapitel 5.1) im vorhinein die Anzahl an Dokumenten und der daraus zu extrahierenden Terme eingegrenzt werden.[88]

85 Eine semantische Umfeldanalyse oder der Abgleich mit einer Personendatenbank wären hier denkbar.
86 Durch Anreicherung des Textes mit entsprechenden Metainformationen wäre die Aussage direkt herleitbar.
87 ANNIE: a Nearly-New Information Extraction System, http://gate.ac.uk/annie/index.jsp
88 [Kushmerick2001], [Zavrel2000]

3.4 Knowledge Extraction

Verbindet man die in Kapitel 3.2 und 3.3 dargestellten Verfahrensansätze und benutzt sie dazu, nicht nur einzelne Wörter oder Dokumente aus einer Textmenge, sondern Zusammenhänge, Strukturen, Aussagen oder Konzepte zu extrahieren, so spricht man von *Knowledge Extraction* bzw. K*nowledge Retrieval*. Solche Strukturen können auch über Dokumentgrenzen hinweg ermittelt werden. So können bspw. in einer Dokumentkollektion häufig vorkommende Wortsequenzen extrahiert werden.[89] Bei häufigem Auftreten solch einer Sequenz könnte diese als "Allgemeinwissen" verstanden werden. Für die Extraktion von Wissen aus unstrukturierten Texten reichen nach U. Hahn lexikonbasierte und statistische Mittel allerdings nicht aus:

> "However, any attempt targeted at the extraction of facts, propositions or even more ambitious evaluative statements from texts using such lexico-statistical methods is doomed to failure. This is where linguistic and knowledge-based approaches to natural language text analysis come in."[90]

Ein Ansatz, mit linguistischen Mitteln Wissen (aus unstrukturierten Texten) zu extrahieren, stellt das Natural Language Processing (**NLP**) dar.[91] Es soll das immanente Wissen eines Textes in maschinenlesbare Form übersetzen. Dabei wird der Text mithilfe von Parsing bzw. Part-of-Speech (**POS**)-Tagging linguistisch zerlegt und in solche Aussagen umgewandelt, mit denen das System umgehen kann. Ein Beispiel für eine Parsing-Anwendung ist das sog. "Natural Language Toolkit".[92]

Erst durch grammatikalische Analyse, d.h. Syntaxanalyse, kann das System erkennen, dass z.B. die Sätze "x schrieb y" und "x ist der Autor von y" gleichbedeutend sind.[93] Hierfür müssen u.a. Subjekt und Objekt erkannt werden. Eine entsprechende Darstellung im System wäre dann "Autor(x,y)". Erst durch eine solche Abbildung in Konstantensymbole (x,y) und Aussagensymbole (Autor) ist die Anwendung von Logik und das Gewinnen neuen Wissens möglich. So könnte das System z.B. aus den Sätzen "A ist Bruder von B" und "B ist Schwester von C" schließen, dass A ein Bruder von C ist. Für diese Herleitungen ist neben der Syntaxanalyse allerdings noch das Erkennen der semantischen Beziehungen von "Autor" und "Schreiben" sowie von "Schwester" und "Bruder" erforderlich (s. Kapitel 4.5).

---

89 [Ahonen-Myka2002], [Ahonen1998]
90 [Hahn1997]
91 Eine Liste an NLP-Anwendungen ist u.a. auf http://opennlp.sourceforge.net/ zu finden.
92 http://nltk.sourceforge.net/
93 [Lin2001]

Genau wie beim Information Extraction kann hier der Einsatz semantischen Metawissens bei der Einordnung der extrahierten Terme behilflich sein. Eine Beispielanwendung für diese Vorgehensweise ist der "TextAnalyst" von Megaputer[94]. Ein vielversprechender Ansatz bei der sog. *Concept Discovery* ist desweiteren die Kombination semantischer Netze mit statistisch orientierten Darstellungsformen wie Clustering (s. Kapitel 5.2), da sie einerseits bestehende Beziehungen der Wörter und zum anderen das Umfeld der Begriffe berücksichtigt.[95] Um diese Strukturen nutzbar zu machen, müssen sie durch Assoziationsregeln (s. Kapitel 4.3) festgehalten und durch einen Indexierungsvorgang (s. Kapitel 4.2) in Maschinensprache übersetzt werden.

Die Extraktion von Wissen aus strukturierten Dokumenten oder aus Knowledgebases stellt sich als einfacher heraus, da dort das Wissen in mehr oder weniger eindeutiger Form abgebildet wird und nur durch entsprechende Abfragen aus dem System extrahiert werden muss. Einfache Zusammenhänge wie z.B "Autor(x,y)" können so zwar bspw. in einer Ontologie abgebildet werden (s. Kapitel 5.5), für das Retrieval solch eines Zusammenhanges bedarf es allerdings wiederum ggf. einer Ausweitung bzw. Umformulierung der in natürlicher Sprache formulierten Suchabfrage (s. Kapitel 6.3).

94 Text Analyst: http://www.megaputer.com/products/tm.php3
95 [Lin2002]

4 Analyse

Das Objekt der Analyse kann das immanente Wissen oder das Metawissen sein. Üblicherweise wird das interne Wissen am externen Wissen gemessen. In diesem Sinne stellt das Metawissen das Messinstrument dar, es kann aber auch selbst analysiert werden (z.B. inwieweit es die Textdokumente adäquat beschreiben kann). Das Ergebnis dieser Analysen stellt wiederum Metawissen dar.

Werden Textmengen untersucht, so besteht das Ergebnis entweder nur aus Begriffen, die in der untersuchten Wortmenge bereits vorhanden sind, oder es werden neue hinzugefügt. Dieses Abgrenzungsmerkmal lässt sich z.B. auf die Unterscheidung der später vorgestellten Methoden Klassifikation (s. Kapitel 5.1) und Clustering (s. Kapitel 5.2) beziehen: Ersteres fügt dem Betrachtungsobjekt einen Klassen- oder Kategorienamen hinzu, während der Cluster nur eine Zusammenstellung vorhandener Informationen darstellt.

Die Analyse wird üblicherweise in "Trainingsphase" und "Anwendungsphase" unterteilt. Zunächst wird das System in einem kontrollierten Lernprozess auf der Basis vorhandener Texte (manuell) angepasst. In der zweiten Phase wird es selbstständig den Rest der Dokumente oder neue Dokumente bearbeiten bzw. analysieren. Man unterscheidet dementsprechend auch zwischen *supervised* und *unsupervised learning*. Auf den Einsatz der Klassifikation bezogen bedeutet dies, dass die supervised learning-Phase aus einer manuellen Anpassung eines vorhandenen Klassifikationssystems besteht und die unsupervised learning-Phase aus der automatischen Einordnung von Wissensobjekten in diese Struktur. Die Trainingsphase wird optimalerweise durch Visualisierung der Zwischenergebnisse unterstützt, um manuelle Änderungen vornehmen zu können.[96] Beim unsupervised learning wird entweder das bestehende Wissen auf das neue Wissen angewandt oder wie z.B. beim Clustering lediglich eine andere Sichtweise auf vorhandene Daten gegeben.

Im Folgenden werden sowohl statistische als auch linguistische Analyseverfahren vorgestellt, die den anschließenden Abbildungsprozess vereinfachen sollen. Gemeinsam dienen sie der Entwicklung einer Systemsprache, die das Suchen in bzw. nach Dokumenten ermöglicht.

96 [Ester2000], [Ankerst2000]

4.1 Normierung

Die Analysemöglichkeiten hängen stark davon ab, in welcher Form das Wissen vorliegt. Bevor eine Analyse über die Textmenge erfolgen kann, sollte das Vokabular in eine einheitliche Form gebracht werden, damit später beim Retrieval nicht alle grammatikalischen Flexionsformen berücksichtigt werden müssen. *Normierung* bedeutet, eine eindeutige Namenskonvention zu schaffen, mit der das System weiterarbeiten kann.[97] Ziel der Normierung ist es, einen allgemeingültigen *Deskriptor* zu finden, der dem Dokument zugeordnet wird und es eindeutig beschreibt. Mit dem entstehenden Hilfskonstrukt, der sog. *Normalform*, wird anstelle des im Text vorkommenden Begriffes weitergearbeitet. Bei diesem Vorgehen tritt prinzipiell ein linguistischer Informationsverlust auf, der aber nur für Analysen über den Gebrauch der Wörter relevant ist und inhaltliche Analysen nicht betrifft.

Folgende Objekte können Gegenstand der Normierung sein:
- Textdokumente: Wörter, Meta-Angaben
- Datenbanken: Tabellennamen, Datenfelder
- XML-Dokumente: Tagnamen, Attributnamen, Namespacedefinitionen

Die Normierung der Datumsangabe bedeutet z.B. die Überführung in ein einheitliches Format.[98] Ortsangaben können ebenfalls normiert werden, indem man sie durch einheitliche Ländercodes o.ä. ersetzt.

Der Normalisierungsprozess für Textdokumente besteht aus folgenden Schritten: Zunächst werden alle gefundenen Zeichenketten (Strings), d.h. Zeichenketten zwischen Leerzeichen oder Satzzeichen, aus den Dokumenten extrahiert und daraus Terme gebildet (term extraction).[99] Mithilfe linguistischer Analyse werden anschliessend alle gefunden Terme einer Normalform zugeordnet. Dies geschieht hauptsächlich durch:
- Lemmatisierung (Grundformerzeugung),
- Derivation (Wortableitung),
- Dekomposition (Kompositazerlegung) und
- Bindestrichergänzung.

Für diese Schritte ist die Verwendung von Metawissen Voraussetzung, d.h. Flexionsformen und Komposita können nur anhand eines Wörterbuches oder eines semantischen Netzes erkannt und auf die Grundform reduziert bzw. zerlegt werden.

---

97 Der Begriff Normierung wird hier synonym zu "Normalisierung" verwendet.

98 Ein Beispiel für Differenzen bzgl. des Datumsformats sind die Weblogformate "W3C Extended Log File Format", "IIS Log File Format" und das "NCSA Common Log File Format". Denkbar ist neben "08/Apr/2001:17:39:04-0800", "2001/04/08 17:39:04" aber auch "der 8. April im Jahre 2001".

99 [Feldman1998], S.4,S.65-73

4.2 Indexierung

Um die Dokumente schneller zugreifbar und durchsuchbar zu machen, werden alle in der Kollektion gefundenen Begriffe z.B. in einer *invertierten Liste* gespeichert. Durch die darin vorgenommene Sortierung kann, verglichen mit der sequentiellen Suche durch die gesamte Kollektion, ein schnellerer Zugriff erfolgen.

Ein alternatives Indexierungsverfahren ist das *Vektorraummodell*[100]. Es wird ein Vektorraum aufgespannt, in dem jeder Begriff, der in der Kollektion vorkommt, eine Dimension darstellt. Allen Dokumenten wird nun ein Vektor zugeordnet, der nur aus Nullen und Einsen besteht, bzw. aus der Anzahl der Vorkommnisse im Text. Nach der Einordnung des Dokumentes in den Vektorraum kann beim Retrieval stellvertretend für die Dokumente nur noch mit den Vektoren gearbeitet werden.

Der Aufbau eines Indexes für eine vorhandene Kollektion besteht aus folgenden grundlegenden Schritten:
1.Extraktion von Indextermen (Information Extraction),
2.Ermittlung formaler Daten und
3.Übernahme in den Index.

In der ersten Phase werden aus den im Text vorliegenden Zeichenfolgen (Strings) geeignete Indexterme ermittelt. Damit sich der Index nicht mit hochfrequenten und daher irrelevanten Termen aufbläht, werden diese durch sog. *Stoppwörteridentifikation*, d.h. durch Vergleich mit einer Stoppwortliste, herausgefiltert. Satz- und Sonderzeichen werden ebenfalls nicht in den Index übernommen.

Bei der Übernahme formaler Daten über die Indexterme wird zunächst mehr oder weniger genau die Stelle vermerkt (Dokumentnummer, Zeilennummer, Wortnummer), wo sich der Begriff befindet. Mithilfe von Syntaxanalyse bzw. des o.g. POS-Tagging lassen sich zusätzlich die Sätze zerlegen und die grammatikalische Rolle der Wörter (Subjekt, Objekt usw.), bestimmen und vermerken.

100 [Salton1983]

Vor der Übernahme in den Index werden die Strings mit einem Vokabular der normierten Wörter abgeglichen und dann anstatt des Strings die Normalform als Indexterme verwendet:

> "Für den Einsatz eines semantischen Netzes zur Indizierung von Dokumenten bedarf es dabei einer hohen begrifflichen Konsistenz zwischen den in den Dokumenten und dem Netz verwendeten Begriffen, die über Text Mining-Verfahren sichergestellt werden kann. In der Praxis optimiert und beschleunigt der vorgestellte Ansatz die Erstellung des Netzes und automatisiert den Vorgang der Dokumentenindizierung."[101]

Zusätzlich zu dieser Normalform können auch noch Begriffe dem Index hinzugefügt werden, die Ähnlichkeiten zu denjenigen aus dem verfügbaren Metawissen aufweisen. So können Klassenbeschreibungen, Begriffe aus Stichwortkatalogen (wie z.B. der Schlagwortnormdatei (SWD)[102]), Lexika, Deskriptorwörterbüchern oder Thesauri (s. Kapitel 5.4) stellvertretend für den im Text vorgefundenen String aufgenommen werden. Dementsprechend wird von *Klassifizierung*, *Verschlagwortung* oder im weiteren Sinne von *Inhalts-erschließung* gesprochen. Die externen Wissensquellen können also das Dokument anreichern und zum besseren Verständnis oder der Einordnung dienen. Auf die hierfür notwendige Ähnlichkeitsanalyse sowie auf mögliche Darstellungsformen unabhängigen Metawissens wird in den folgenden Kapiteln eingegangen.

Ein Beispiel für die Verwendung von Metawissen bei der Indexierung stellt das Milos-Projekt[103] dar. Hier wurden die Thesaurirelationen der Schlagwort-normdatei mit in die Indexierung einbezogen. In Retrievaltests wurde gezeigt, dass mit dieser Vorgehensweisen der *Recall* wesentlich erhöht werden kann und die *Precision* nur geringfügig verschlechtert wird.[104] Ein weiteres Beispiel stellt J. Gonzalos Ergänzung des Indexes mit Synonymieeinträgen aus dem semantischen Netz WordNet[105] dar. Hier konnten in Bezug auf die Testmenge im Vergleich zu Saltons Vektormodell (SMART) sowohl Recall als auch Precision verbessert werden.[106]

---

101 [Beier2003]
102 http://www.ddb.de/professionell/swd.htm
103 http://www.ub.uni-duesseldorf.de/projekte/milos/mil_home [Lepsky1997]
104 [Gödert1998]
105 WordNet: http://www.cogsci.princeton.edu/~wn
GermaNet (das deutsche Äquivalent zu WordNet) http://www.sfs.nphil.uni-tuebingen.de/lsd/
106 [Gonzalo1998]

In den Index wird also die Normalform, ein entsprechender Lexikoneintrag, Deskriptor und/oder eine Klassennotation stellvertretend für den im Text vorgefundenen String aufgenommen. Dadurch kann bei der Analyse oder Suche nur noch mit dem Index bzw. der Abbildung des Dokumentes in die Dokumentationseinheit (**DE**) gearbeitet und der Volltext (die Dokumentarische Bezugseinheit (**DBE**)) nur noch als Referenz verwendet werden.[107]

Statt der Aufnahme verwandter Begriffe in den Index kann diese Zuordnung aber auch erst bei der Suche erfolgen, indem der Suchstring um im Metawissen vorgefundene verwandte Begriffe erweitert wird (semantische Umfeldsuche).[108] Die Vor- und Nachteile dieser beiden Vorgehensweisen können hier nicht dargestellt werden. In beiden Fällen werden zwei Indizes aufgebaut: ein Index über die in den Texten gefundenen Wörter und einer, der alle Begriffe des verwendeten Systemlexikons enthält. Bei Fehlen von background knowledge und Normierungsverfahren müssen alle in der Gesamtheit der Texte gefundenen Strings in den Index aufgenommen werden.

107 [Stock2000a], S.123
108 [Lepsky1998]

4.3 Ähnlichkeit

Ähnlichkeit stellt eine Relation von Objekten dar, die umgangssprachlich auch als "Verwandtschaft" oder "Übereinstimmung" bezeichnet wird.[109] Im TM werden hauptsächlich Ähnlichkeiten von Dokumenten bzw. von Begriffen (Suchanfragen) zu Dokumenten betrachtet. Bei solchen Wissensobjekten eine semantische Ähnlichkeitsrelation aufzustellen, bedeutet, den Gebrauch des Vokabulars zu betrachten. Ziel der Ähnlichkeitsanalyse ist bspw. die Einordnung des Wissensobjektes in eine Klasse oder einen Cluster. Gleichzeitig mit der Einordnung geschieht eine Abgrenzung zu anderen Klassen oder Clustern. Die nachfolgend vorgestellten Verfahren können sowohl bei der Indexierung als auch bei der Suche nach ähnlichen Dokumenten behilflich sein.

Ähnlichkeit wird mithilfe von Ranking-Algorithmen quantifiziert und meist in Form einer reellen Zahl zwischen 0 und 1 ausgedrückt, wobei 0 die Abwesenheit von Ähnlichkeit und 1 vollkommene Ähnlichkeit (Identität) bedeutet. Das gemeinsame Auftreten von Begriffen deutet auf eine Ähnlichkeit der Begriffe oder zumindest einen gemeinsamen Kontext hin. Dementsprechend wird nicht von Ähnlichkeit, sondern von *Zusammenhang*, *Korrelation*, oder *Assoziation* gesprochen, welches das gemeinsame Auftreten allgemeiner beschreibt.

Die Ähnlichkeit kann z.B. durch den Jaccard-Sneath-Koeffizienten[110] oder durch eine Assoziationsregel R : (W → w) ausgedrückt werden, bei der W eine Menge an Schlüsselwörtern und w ein Schlüsselwort ist, das mit W assoziiert wird. Die Gültigkeit dieser Regel für eine Kollektion T kann durch Support- und Confidence-Werte beschrieben werden:

"Support: $S(R,T)=\left|[W\cup\{\mathrm{w}\}]\right|$, Confidence: $C(R,T)=\frac{\left|[W\cup\{\mathrm{w}\}]\right|}{\left|[W]\right|}$ "[111]

> "Confidence is the proportion of texts that have X AND Y in relation to the number of texts that have only X, and support is the proportion of texts that have X AND Y in relation to all texts in the collection. Confidence works like the conditional probability (if X is present, so there is a certain probability of Y being present too)."[112]

109 Aus: http://www.wissen.de und Wahrig Deutsches Wörterbuch.
110 [Stock2000a], S.141
111 [Besançon1998]
112 [Loh2003], S.360

S. Loh et al setzen den Supportwert also in Relation zur Gesamtkollektion. R. Besançons Experimente zeigten, dass die Assoziationsregeln, wenn sie nicht über den Index bzw. die keywords, sondern über den Volltext erfolgen, zwar statistisch signifikant sein können, aber teilweise uninterpretierbare oder unsinnige Ergebnisse zutage bringen können.[113]

Ähnlichkeiten können mithilfe verschiedener Verfahren ermittelt werden, von denen im Folgenden drei häufig verwendete betrachtet werden.

Auf Basis des invertierten Indexes gibt die Anzahl der in beiden Dokumenten vorkommenden Begriffe die Ähnlichkeit an.

Beim Vektormodell gibt der Winkel zwischen den Vektoren oder das Vektorprodukt der Vektoren die Ähnlichkeit der beiden Texte an. Existiert ein Begriff in einem Dokument, im anderen jedoch nicht, so ergibt sich durch die Multiplikation mit 0 ein "kleinerer" Ergebnisvektor und somit eine geringere Ähnlichkeit. Mithilfe dieses Verfahrens können auch Ähnlichkeiten von Wörtern ermittelt werden. Auf Basis dieser Ergebnisse sind dann Regeln zur Wahrscheinlichkeit des Auftretens anderer Wörter in einem bestimmten Wortumfeld (window) herleitbar.[114]

Die Hyperlinkanalyse betrachtet die Verknüpfungen der Dokumente als Indiz für Ähnlichkeit. Sie basiert auf der in den 1950er Jahren von Eugene Garfield[115] entwickelten Zitationsanalyse, die bei dem sog. "Science Citation Index" die Grundlage darstellt. Der Thomson-Konzern[116] benutzt diesen Index in erweiterter Form für etwa 6000 Zeitschriften. Hintergrund für dieses Verfahren ist die Annahme, dass ein Link bzw. Zitat durch intellektuelle Eingabe erzeugt wird und der Vertiefung, Verdeutlichung und/oder Fortführung des Themas dient. Ein von einer anderen Seite eingehender Link wird dabei als *Zitation* und ein ausgehender Link als *Referenz* bezeichnet. Links können neben ihrer Richtung auch noch in andere Kategorien eingeteilt werden, z.B. in "kritisierend", "zustimmend", "erläuternd", "verbindend" oder "fragend"[117]. Diese Zitationsarten finden z.B. bei "Shepard's Citation Service"[118] Anwendung und können dort zur Eingrenzung der Suche herangezogen werden.

113 Das gewählte Beispiel {wall} → street zeigt, dass die Verwendung von Data Mining-Techniken ohne semantische Hilfsmittel nicht aussagekräftige oder unsinnige Ergebnisse produzieren kann.
114 [Ahonen1997]
115 http://www.garfield.library.upenn.edu
116 http://www.thomson.com
117 [Stock2000a], S.303
118 http://www.lexisnexis.com/shepards/

Anhand der Struktur eines Links lässt sich auf die Spezifität oder Art des verlinkten Dokumentes schließen. So beschreibt der Link http://www.search.com/search.cgi?text+knowledge bspw. ein virtuelles oder dynamisches, d.h. nicht real existierendes Dokument, während http://www.search.com/help/index.html ein "reales", statisches Dokument und http://www.search.com/help/index.html#sort eine genaue Position in diesem Dokument darstellt.

Vorhandenes lineares (nicht verlinktes oder verlinkendes) Textmaterial lässt sich auch im Nachhinein noch mit Links anreichern,[119] die dann in die Linkanalyse übernommen werden können. Auch hierbei können verschiedene Linktypen spezifiziert werden.[120]

Die Ähnlichkeitsanalyse kann auch über mehrere Instanzen erfolgen, d.h. wenn A zu B ähnlich ist und B zu C, dann weist auch A zu C eine Ähnlichkeitsbeziehung auf. So kann Ähnlichkeit gewissermaßen "vererbt" bzw. durch mehrere Instanzen analysiert werden. Angewendet auf die Zitationsanalyse bedeutet dies, dass zwei unterschiedliche Dokumente, die beide dasselbe Dokument zitieren, ebenso eine Ähnlichkeitsbeziehung aufweisen wie zwei Dokumente, die in demselben Dokument zitiert werden. Diese Ähnlichkeitsbeziehungen werden mit der *Kozitationsanalyse* abgedeckt.

Ein Beispiel für die Anwendung mehrerer Verfahren der Ähnlichkeitsanalyse stellt der "ResearchIndex" dar, der mithilfe des sog. "Autonomous Citation Indexing"-Algorithmus[121] erstellt wurde und bei der Suchmaschine Citeseer[122] Anwendung findet. Hier werden zusätzlich zu der Auflistung von Zitationen, Referenzen und Kozitationen für ein Dokument noch Ähnlichkeitswerte für Vergleiche auf Satz- und Volltextebene gegeben. Citeseer zeigt, dass es nicht sinnvoll ist, die Ähnlichkeitsbeziehungen zweier Dokumente in einem verallgemeinerten Ähnlichkeitswert darzustellen, sondern separat aufzulisten.

Die Ähnlichkeitsanalyse von Dokumenten kann durch Heranziehen von Metawissen wie z.B. Wörterbüchern erweitert werden. So können Dokumente bspw. als ähnlich eingestuft werden, in denen synonyme Begriffe benutzt werden. Die Ähnlichkeit im Metawissen wird somit auf die Dokumente übertragen.

---

119 [Myka1992], [Myka1995], [Myka1996a]
120 page links, hierarchical links, similarity links, syntactical links, virtual links, vgl. [Myka1996c]
121 [Lawrence1999]
122 http://citeseer.org , http://citeseer.ist.psu.edu, [Giles1998], [Bollacker1998]
Visualisierung von Dokumentrelationen in CiteSeer: http://www.pmbrowser.info/citeseer.html

4.4 Relevanz

Basierend auf der Ähnlichkeitsanalyse können die für eine Suchanfrage *relevantesten* Texte ermittelt werden. Die Suchanfrage stellt dabei im Retrievalsystem ein mit einem Dokument vergleichbares Objekt dar, das auf Ähnlichkeit untersucht wird. Die Dokumente, die Übereinstimmungen mit den meisten Begriffen aufweisen können, werden nach oben sortiert. Zusätzlich kann aber noch eine Gewichtung der gewählten Suchbegriffe vorgenommen werden, und zwar durch sog. *Relevanz-* oder *Gewichtungsfaktoren*.

> "Eine Möglichkeit zur Berechnung von Deskriptorgewichten basiert auf der Annahme, die Bedeutung eines Begriffes sei proportional zur Häufigkeit des Begriffes im Dokument und umgekehrt proportional zur Gesamtanzahl der Dokumente, denen der Begriff als Deskriptor zugeordnet ist."[123]

Aus diesen Überlegungen resultieren die beiden Kennzahlen WDF (within document frequency) und IDF (inverted document frequency). WDF setzt die Häufigkeit der Begriffe TF (term frequency) in Relation zur Begriffsanzahl des Dokumentes und IDF setzt sie in Beziehung zum Auftreten des Begriffes in der gesamten Kollektion. Üblicherweise werden beide Kennzahlen in einer Gewichtungsfunktion zusammengefasst, d.h. bei der Bestimmung des Gewichtungswertes kommen beide Werte zum Einsatz. Der Überbewertung von nicht trennscharfen Begriffen wird damit entgegengewirkt.

Bei Googles Rankingalgorithmus "PageRank"[124] dient die Linkanalyse als Hauptindikator bei der Bewertung einer Webseite. Google geht von der Annahme aus, dass häufig verlinkte, also populäre Dokumente relevanter sind als weniger verlinkte. So werden bspw. zwei einer Suchanfrage gleich ähnliche, aber unterschiedlich populäre Dokumente dementsprechend aufgelistet. Google ist desweiteren ein Beispiel dafür, dass auch Relevanz vererbt werden kann. So steigen Seiten in ihrer Popularität, die von Seiten verlinkt werden, die selber häufig verlinkt wurden.

Je nach formeller Beschaffenheit des vorliegenden Materials können auch formelle Gesichtspunkte das Ranking beeinflussen. So vergeben E. Desmontils und C. Jacquin bei der Indexierung von HTML-Dokumenten bestimmten HTML-Tags und verwendeten Schriftgrössen unterschiedliche Wertungen.[125]

123 [Salton1983]
124 eingesetzt bei http://www.google.com (Autoren: Brin und Page)
125 [Desmontils2001]

5 Abbildung

Die Ergebnisse der Ähnlichkeits- sowie der Relevanzanalyse können in den Index der Dokumentkollektion übernommen oder separat als Metawissen abgespeichert werden und für weitere Dokumente genutzt werden. Für die Abbildung des Metawissens existieren verschiedene Darstellungsformen, von denen einige häufig verwendete im Folgenden vorgestellt werden. Werden sie im Zusammenspiel mit o.g. Analyseverfahren nicht bloß zur Indexierung verwendet, sondern dazu eingesetzt, die Dokumente durch eine Systemsprache abzulösen, so spricht man von sog. *Wissenskonversion*. Je nach Qualität der Darstellungsmethoden kann dieses Ziel mehr oder weniger erreicht werden. Die hier vorgestellten Ansätze nehmen ihrer Reihenfolge entsprechend an Ausdrucksstärke und Flexibilität, d.h. in ihrer Fähigkeit, Wissen darzustellen, zu. Ontologien (s. Kapitel 5.5.) scheinen derzeit das beste Mittel zu sein, die Vielfalt an semantischen Beziehungen von Wissensobjekten darzustellen.

Mit der Frage, welche Darstellungsformen allgemein als Wissensstrukturen bezeichnet werden können, setzt sich u.a. die KR, Inc[126] auseinander. In dem Artikel "What is a Knowledge Representation?"[127] werden die verschiedenen Interpretationsmöglichkeiten einer Wissensrepräsentation aufgezeigt.
Als sog. "knowledge representation technologies" werden dort "basic representation tools like logic, Prolog, Lisp, predicates, rules, frames, semantic nets, etc." genannt. Diese Technologien stellen Sprachen bzw. Systeme dar, mit denen Knowledgebases bzw. sog. *Knowledge Representation Systeme* (KRS) entworfen werden. Sie werden auch *Knowledge Representation Languages* (KRL)[128] oder *Knowledge Modelling Languages* (KML) genannt.
In den damit geschaffenen Systemen soll das Wissen eingebettet und beschrieben werden.

126 Principles of Knowledge Representation and Reasoning, Incorporated, http://www.kr.org
127 [Davis1993]
128 [Bobrov1977]

Die nachfolgend vorgestellten Verfahren können als *Wissenssprachen* bzw. *Wissenssysteme* angesehen werden, da sie die eingangs erwähnten Kriterien zur Darstellung von Wissen erfüllen: sie sind veränderbar, weisen eine Struktur auf, werden der Kontextbezogenheit eines Faches, Themas oder Begriffes gerecht und dienen als Mess-, Beurteilungs- und Einordnungsinstrument für neue Informationen.

M. Lusti unterscheidet zwischen relationsorientierter und objektorientierter Wissensdarstellung.[129] Während bei relationsorientierten Ansätzen mit Aussagenlogik bzw. Beschreibungslogik (description logics) gearbeitet wird, werden bei den objektorientierten Ansätzen Graphen bzw. semantische Netze eingesetzt. Erstere stellen die Grundlage für sog. *Knowledge Representation and Reasoning* (KRR)-Systeme dar, in denen neben der Darstellung von Wissen die Möglichkeit gegeben wird, mit diesem Wissen zu arbeiten und neue Erkenntnisse herzuleiten. Da dies allerdings über die Darstellung von Wissen hinausgeht und bereits die Verwendung dieses Wissens beschreibt, wird auf diese Systeme in diesem Kapitel nicht eingegangen. Der Schwerpunkt wird daher auf die semantischen Verfahren gelegt.

129 [Lusti1990], S.225,S.249

5.1 Klassifikation

Unter *Klassifikation* wird zum Einen der Prozess verstanden, bei dem Dokumenten Klassenbeschreibungen (*classifier*) zugeordnet werden, die deren Thematik ausdrücken sollen.[130] Zum Anderen ist mit Klassifikation das semantische Netz gemeint, das bei der Indexierung verwendetet wird.[131]
Die verwendeten Klassenbeschreibungen können hierarchisch oder sequentiell angeordnet werden, wobei hier eine Mischung möglich ist.

Manuell erstellte Klassifikationen stellen sog. *Domänenwissen* dar, wenn sie ein klar umrissenes Themenfeld beschreiben. So sind ISIC, NACE[132] und eCl@ss[133] bekannte Produktklassifikationen. Beispiele für eher allgemeine Klassifikationen sind die "Dewey Dezimalklassifikation" (DDC)[134] und die davon abgeleitete multilinguale "Universal Decimal Classification" (UDC)[135].
Neben diesen nicht frei verfügbaren Systemen existieren im WWW einige Klassifikationen, die sich auch in eigene Systeme integrieren lassen.[136]
Yahoo[137] und Dmoz[138] stellen dabei die wohl bekanntesten manuell gepflegten Klassifikationssysteme des WWW dar. Dmoz ist durch die Integration in Google[139] bekannt geworden und ermöglicht es jedem Interessierten, an der Klassifikation mitzuarbeiten.

Diese manuell erstellten Klassifikationen lassen sich zur Einordnung von Dokumenten bzw. zur Vergabe von Metainformationen verwenden. Dies kann ebenfalls manuell vorgenommen werden oder automatisch erfolgen. Bei der automatischen Klassifikation eines Dokumentes werden Ähnlichkeitswerte zu bereits einsortierten Dokumenten ermittelt und bei Überschreitung eines Grenzwertes dieselbe Klassenbezeichnung vergeben. Werden Ähnlichkeiten zu verschiedenen Klassen ermittelt, so muss das Dokument in die allgemeinere (höhere) Klasse oder in mehrere Unterklassen einsortiert werden. Ersteres hätte allerdings einen gewissen Informationsverlust zur Folge. E. Riloff und W. Lehnert zeigen, wie die Relevanzanalyse zur Klassifikation herangezogen werden kann.[140]

---

130 Die Begriffe Indexierung und Klassifizierung wären diesbezüglich passender.
131 Werden das Klassensystem im Umfang sowie in der Hierarchietiefe stark eingegrenzt, wird eher von "Kategorisierung" gesprochen.
132 http://www.fifoost.org/database/nace/
133 http://www.eclass.de
134 http://www.oclc.org/dewey/ , http://www.ddc-deutsch.de
135 http://www.udcc.org
136 [Labrou1999], [Mladenic1998]
137 http://www.yahoo.com, http://www.yahoo.de
138 http://www.dmoz.org
139 http://www.google.de/dirhp
140 [Riloff1994]

Neben diesen bereits vorgestellten Verfahren können auch Verfahren der AI wie z.B. neuronale Netze und Entscheidungsbäume (*decision trees*) bei der Klassifikation zum Einsatz kommen.

Im TM kann ein neuronales Netz für die Bestimmung der passenden Klassenbeschreibung genutzt werden. Dabei werden werden als Ein- und Ausgabewerte der Neuronen die Gewichtungsfaktoren verwendet, die so lange angepasst werden, bis unter Betrachtung der Ähnlichkeit zu der Klasse oder Cluster die optimale Klasse gefunden wird.[141] Für das Trainieren des neuronalen Netzes sind bestehende Ähnlichkeitswerte für eine repräsentative Menge von Dokumenten Voraussetzung.[142] Bei der Nutzung von neuronalen Netzen zur internen Darstellung von Texten ergibt sich das Problem, dass die Textobjekte durch Zahlenwerte dargestellt werden müssen. Denkbar wäre hier die Nutzung der bereits bestehenden Vektorenwerte und statistischen Ergebnisse der Ähnlichkeitsanalyse.

Entscheidungsbäume werden dazu benutzt, anhand von dort festgelegten Regeln eine Aktion wie z.B. die Einordnung in eine Klasse zu vollziehen. Entlang des Entscheidungsbaumes wird gemäß dieses Regelwerkes bspw. die Entscheidung "irrelevant – relevant" oder "ähnlich – verschieden" getroffen. Solch eine Entscheidung muss im Sinne der Indexierung für die Vergabe eines classifiers für ein Dokument getroffen werden. Die decision trees werden mit Data Mining Verfahren solange abgeändert, bis die classifier ein möglichst hohes Abgrenzungspotential, d.h. eine hohe Relevanz erreichen.[143]

Beide Verfahren stellen sich in der Anwendung zwar als praktikabel, aber durch die ständige Anpassung der Neuronen bzw. decision trees für den Benutzer als uneinsehbar ("*black box*") oder zu kompliziert heraus:

> "Decision trees were left out, because most concepts are not exclusive of one class and since the combination of concepts could form complex rules with many attributes, making difficult to understand or to explain the rules used in the decision process. Other alternative was to use a neural net. However, this kind of decision model cannot explain how the decision was obtained."[144]

Auf die rudimentären Verfahrensansätze, die Entwicklung eines neuronalen Netzes visuell zu unterstützen,[145] kann hier nicht weiter eingegangen werden.

---

141 [Dasigi1996]
142 [Callan2003]
143 [Quinlan1986]
144 [Loh2003]
145 "Visualization and interpretation of neural feedforward networks is possible not only for two but even for higher dimensions. The neural network can be understood better and does not look like a black box." aus: [Bernatzki1996]

Ein bestehendes Klassensystem kann mithilfe von **Bayes**-Algorithmen erweitert werden, indem Wahrscheinlichkeitswerte für Begriffe hinzugefügt werden, die in Dokumenten dieser Klasse vorkommen.[146] Das Vokabular der Klasse (classifier ci) wird so lange mit einer Kollektion an Begriffen w1-wm ("bag of words") erweitert, bis die Wahrscheinlichkeit des Auftretens der gewählten Begriffe maximiert wird.

> "Modeled as generating a bag of words for a document in a given category by repeatedly sampling with replacement from a vocabulary V={w1, w2,...wm} based on the probabilities P(wj | ci)."[147]

Klassensysteme decken durch ihre hierarchische Struktur eine Ober-Unterbegriff–Relation und durch Verweise auf andere Klassen eine nicht näher definierte Assoziationsrelation ab. Die Vielfalt an Assoziationsrelationen[148], die bspw. durch die o.g. Zitationsanalyse ermittelt werden können, lässt sich damit nicht (explizit) darstellen.

Die Anwendung des Klassensystems auf die Dokumente beim Indexieren bietet folgende Möglichkeiten:

- Die Klassenbeschreibung liefert eine eindeutige Themenbezeichnung und dient damit der Orientierung.
- Die Suche nach thematisch ähnlichen Texten wird möglich. Sie lässt sich auf die den Suchtermen ähnlichsten Klassen beschränken und ermöglicht dadurch ein effizientes Ausfiltern von Texten.[149]
- Die Suche kann durch die Hierarchie eingegrenzt oder ausgeweitet werden, was mit der im Data Mining verwendeten "Drilldown"-Technik vergleichbar ist.

146 [Langley1992]
147 [Mooney2003]
148 [Stock2000a], S.62
149 [Lanquillon2001]

5.2 Clustering

Die strikte hierarchische Struktur eines Klassensystems und somit die damit verbundene Problematik der eindeutigen Klassifizierung bei überschneidenden Thematiken kann mithilfe von Clustern umgangen werden. Die Abbildung hierarchischer Strukturen ist mit Clustern allerdings ebenfalls möglich[150] und im Gegensatz zum Klassensystem können sich Cluster überschneiden. Während Clustering im Data Mining häufig zur Darstellung und Abgrenzung komplexer numerischer natur- sowie wirtschaftswissenschaftlicher Datenräume verwendet wird und dort auch seinen Ursprung hat, kann es im TM der Zusammenfassung von Begriffen, Dokumenten und Themen- oder ganzer Wissenschaftsfelder dienen. Neben Clustern, die aus Ansammlungen von Textdokumenten (bzw. den Dokumentvektoren) bestehen, können auch Begriffscluster gebildet werden.[151] Solche "Wortwolken", von P. Pantel und D. Jin auch "committees" genannt, können zur Klärung von Begriffen[152] oder Erweiterung von Suchanfragen[153] beitragen.

Beim Clustering können überwiegend dieselben Algorithmen wie beim Klassifikationsverfahren verwendet werden, allerdings wird keine Klassenbeschreibung vergeben, sondern lediglich das Zentrum sowie die Grenzen eines Clusters definiert. Durch die Ähnlichkeitsanalyse von umfangreichen wissenschaftlichen Dokumentkollektionen können so gesamte Wissens- oder Forschungscluster gebildet werden. Beim Clustering werden Ähnlichkeitsbeziehungen sowie Unterscheidungsmerkmale herausgearbeitet, anhand derer die Dokumente in zusammengehörige Segmente eingeteilt werden. Zum Ermitteln dieser Cluster werden Charakterisierungs- und Differenzierungsregeln benötigt. Durch Erstere werden die Haupteigenschaften des Clusters erfasst und durch Letztere die Unterschiede zwischen den Clustern beschrieben.

150 [Ding2002]
151 z.B. mithilfe der Textwortmethode, vgl. [Stock2000b]
152 [Pantel2002]
153 [Baeza-Yates1999]

In Anlehnung an M. Ester kann Clustering in einige grundsätzliche Methoden unterschieden werden, die alle eine Distanzfunktion für die Ähnlichkeitsanalyse verwenden.

Bei den partitionierenden Verfahren wird neben der Distanzfunktion eine Anzahl an Clustern vorgegeben, in denen die Distanz zwischen den Dokumenten minimiert wird. So bilden sich Schwerpunkte der Cluster heraus:

> "Each cluster is represented by the gravity center of the cluster (k-means) or by one of the objects of the cluster located near its center (k-medoid)."[154]

Bei den hierarchischen Verfahren wird entweder von der Gesamttextmenge als ein Cluster ausgegangen und dieser in kleinere Untercluster unterteilt ("top-down") oder jedes Dokument als einzelnes Cluster betrachtet und diese zu größeren Clustern zusammengeführt ("down-top").

Die dritte Möglichkeit stellen die dichtebasierten Verfahren dar, in der ein Maximalabstand (Minimaldichte) definiert wird und alle Dokumente, die dieses Kriterium erfüllen, zum Cluster dazugezählt werden.

Zwei häufig verwendete Techniken dieser Zuordnung von Dokumentvektoren zu Clustern sind das **Rocchio** Verfahren und das "**k-nearest neighbor**"-Verfahren (**kNN**). Beide verwenden durch WDF-IDF gebildete Vektoren. Das Rocchio Verfahren ist ein Vektormodellverfahren, bei dem ein bestehendes Klassensystem Voraussetzung ist. Für jede Klasse wird ein Durchschnittsvektor ermittelt, der *Zentroidvektor* oder *Prototyp* genannt wird. Für das Einordnen eines Objektes in eine dieser Klassen ist nun nicht mehr notwendig, Ähnlichkeiten zu allen Dokumenten der Klasse zu ermitteln, sondern nur noch die Ähnlichkeit zu dem Zentroidvektor. Dieser sollte allerdings nach Einsortieren von neuen Dokumente neu ermittelt werden, da er nach dazugekommenen Dokumenten nicht mehr für die ganze Klasse repräsentativ ist. Bei dem kNN-Verfahren wird im Gegensatz zum Rocchioverfahren nicht ein Zentroidvektor zur Ähnlichkeitsbestimmung verwendet, sondern die Ähnlichkeit zu jedem Dokument aus der Testmenge. Die Klasse wird nicht um den Zentroidvektor gebildet, sondern um eine Gruppe an sehr ähnlichen Dokumenten. Es wird also nicht ein Vektor, sondern eine (meist ungerade) Zahl k an Vektoren verwendet.

154 [Ester1998]

Beim Vergleich der verfügbaren Cluster- bzw. Klassifikationsalgorithmen betrachtet Y. Yang die vergebenen Klassenbeschreibungen und bewertet deren Recall- und Precisionwerte in Bezug auf die von ihnen klassifizierten Dokumente.[155] Desweiteren bezieht er fehlerhafte Klassifizierung sowie übernommene Ausreisser mit in die Berechnungen ein.[156] Er kommt zu dem Ergebnis, dass das kNN-Verfahren diesbezüglich ein gutes Mittel darstellt:

> "The impressive Performance of kNN is rather surprising given that the method is quite simple.[...] Rocchio has a relatively poor performance compared to other learning methods.[...] This suggests that Rocchio may not be a good choice (although commonly used)."[157]

Auch D. Michie et al kommen zu diesem Ergebnis.[158] Die Verfahren Decision Tree und Naive Bayes bewegen sich in einem ähnlichen Wertespektrum wie Rocchio.

Beim Clustering müssen im Vorhinein entweder Grenzwerte bezüglich der Dichte bzw. Ausdehnung oder eine feste Anzahl an Hierarchieebenen definiert werden. Ähnliche Designüberlegungen müssen auch für Klassifikationssysteme getroffen werden. Die Klassen- bzw. Clustergröße beeinflusst die Qualität der darauf aufbauenden Ähnlichkeitsanalyse. So liefert z.B. die Erweiterung der Clustergrenzen eine grössere Menge an scheinbar ähnlichen Dokumenten.

Wie beim Klassifikationssystem können die Cluster als Hilfsmittel der Ortung relevanter Texte dienen, indem sie irrelevante Dokumente ausgrenzen. Cluster sind desweiteren dazu geeignet, zwei- oder dreidimensionale "Wolken" von Dokumenten zu visualisieren. Zur Darstellung solcher Cluster können sog. "Self Organising Maps" (SOMs)[159], die Unified Distance Matrix (U-Matrix)[160] sowie weitere Verfahren[161] eingesetzt werden. Das als "document exploration tool" bezeichnete WEBSOM wendet erstgenanntes Verfahren auf die Darstellung von Internetpräsenzen an.[162] Internet-Suchmaschinen wie Kartoo[163], Vivisimo[164] und Objectssearch[165] benutzen ebenfalls Clustering zur thematischen Strukturierung von Suchtreffern. Die dort verwendeten Verfahren konnten im Rahmen dieser Arbeit allerdings nicht ermittelt werden.

---

155 [Yang1997], S.69-90
156 ebd, S.4-5
157 ebd. S.8
158 "Although this method did very well on the whole, as expected it was slowest of all for the very large datasets. However, it is known (Hart, 1968) that substantial time saving can be effected [...]" aus: [Michie1994]
159 [Kohonen2001]
160 [Ultsch2003]
161 [König1998]
162 http://websom.hut.fi/websom/, [Honkela1997]
163 http://kartoo.com
164 http://vivisimo.com
165 http://www.objectssearch.com/de/

Da die Cluster nur eine allgemeine Zugehörigkeit von Dokumenten zueinander definieren, decken sie die Vielfalt an semantischen Beziehungen nicht ab. Eine Verbesserung der Clusterbildung mithilfe von Ontologien[166] oder anderer semantischer Netze ist zwar möglich, doch die anschliessende Darstellung der Cluster lässt diese außer Acht. Neben den Clustern oder Klassen gibt es andere Modelle, die besser die semantischen Beziehungen der Begriffe bzw. Dokumente berücksichtigen können und damit im Hinblick auf die möglichen Beziehungen von Wörtern flexiblere Instrumente darstellen.

5.3 Topic Maps

Mithilfe von sog. *Topic Maps*[167] können beliebig viele eigene Relationen zwischen Wörtern, Wortgruppen o.ä. definiert werden.[168] Üblicherweise werden Topic Maps aber aufgrund ihrer Offenheit und eher grafischen Orientierung primär zur Veranschaulichung in Form sog. *Mind Maps* verwendet, um Präsentationen von Ideen oder Projekten bzw. Zusammenhängen zu unterstützen. Ein Beispiel für eine Anwendung von Topic Maps in diesem Sinne stellt der sog. "MindManager" von MindJet[169] dar. P. A. Kirschner et al beschreiben die Vielfältigkeit an Einsatzbereichen für derartige Darstellungsmittel.[170] Die definierten Beziehungen sind allerdings frei wählbar und bislang nur in geringem Masse standardisiert (S. Kapitel 6.3). Aus diesem Grund können sie zu einer übergreifenden Strukturierung von Dokumenten bzw. der Suche nach Wissen nur begrenzt eingesetzt werden.

166 [Hotho2002], [Hoto2003a], [Hotho2003b]
167 ISO-Standard für Topic Maps: ISO13250
168 Prinzipiell lassen sich mit Topic Maps auch andere Objekte wie z.B. Grafiken oder Tabellen verknüpfen.
169 MindManager von MindMap: http://www.mindjet.com
170 [Kirschner2003]

5.4 Thesaurus

Mit einem *Thesaurus*[171] können im Vergleich zu einer Topic Map eine Reihe an fest definierten semantischen Relationen verwendet werden. Der Thesaurus stellt ein Ordnungssystem für Begriffe dar und baut auf einer sog. *Taxonomie* auf. Die Taxonomie ist eine hierarchische Anordnung von Begriffen und daher vergleichbar mit einem Klassifikationssystem. In einer Taxonomie besitzt jeder Begriff nur zu seinem Oberbegriff und zu seinen Unterbegriffen eine Beziehung. Erweitert wird dieses Modell, indem beim Thesaurus die Ähnlichkeitsrelation von Begriffen (*Deskriptoren*) näher definiert wird. So können in einem Thesaurus eine Assoziationsrelation und sog. *Vorzugsbenennungen* (Deskriptor – Nichtdeskriptor) in Bezug auf synonyme Begriffe definiert werden. Durch Letztere wird eine sog. *terminologische Kontrolle* ermöglicht.[172]
Das nachstehende Beispiel eines Thesauruseintrages wurde dem Rechercheinstrument der Deutschen Bibliothek entnommen:

> "Deskriptor: Neuronales Netz
> BF (benutzt für): Neuralnetz, Neurales Netzwerk, Neurales Netz, Neural network, KNN, Künstliches Neuronales Netz, ANN, Artificial neural network
> OB (Oberbegriff) : Soft Computing
> VB (verwandter Begriff): Neurocomputer, Nervennetz/Modell, Konnektionistisches Netz"[173]

Auch die Synonymie eines zusammengesetzen Begriffs zu den Teilbegriffen kann abgebildet werden (BK: benutze Kombination, KB: Kombinationsbegriff).[174]

Ein Thesaurus kann die hierarchische Struktur eines Klassifikationssystems übernehmen. So werden z.B. beim Standard Thesaurus Wirtschaft[175] die Notationen des o.g. Klassifikationssystemes NACE verwendet. Im Gegensatz zum Klassifizierungssystem oder zur Taxonomie bietet der Thesaurus auch die Möglichkeit, *Polyhierarchien* zu erstellen, d.h. die Zuordnung eines Begriffes zu mehreren Oberbegriffen vorzunehmen. Ein Beispiel für einen von der Allgemeinheit gepflegten Thesaurus stellt der OpenThesaurus[176] dar. Er kann bspw. in Textverarbeitungsprogramme integriert und von jedem Interessierten erweitert bzw. verändert werden.

171 DIN 1463/1: Thesaurus: "[...] geordnete Zusammenstellung von Begriffen und ihren vorwiegend natürlichsprachlichen Beziehungen [...]"
172 [Stock2000a], S.76
173 aus: http://www.ddb.de, http://z3950gw.dbf.ddb.de Schlagwort 4226127-2
174 [Stock2000a], S.79
175 http://www.gbi.de/_de/thesaurus/
176 http://www.openthesaurus.de, http://openthesaurus.sf.net, http://thesaurus.kdenews.org

5.5 Ontologien

Durch sog. *Ontologien*[177] wird dieses Konzept verallgemeinert, indem in ihnen neben semantischen Relationen beliebige andere definiert werden können. Neben den Relationen eines Thesaurus gibt es eine Reihe an weiteren vorgegebenen Relationen. Die sog. "Standard Upper Ontology working group" (SUO WG)[178] arbeitet derzeit an einem Standard, der den allgemeinen Aufbau von Ontologien beschreibt. Die dabei entworfene sog. "Suggested Upper Merged Ontology" (SUMO)[179] beschreibt eine abstrakte Ontologie, die als Basis für domänenspezifische Ontologien eingesetzt werden kann.

Eine grundlegende Eigenschaft von Ontologien ist die Möglichkeit, *Klassen* ("Konzepte"), *Instanzen* ("Objekte") und *Attribute* ("Eigenschaften")[180] darzustellen sowie diese mit vorgegebenen oder selbst definierten Relationen zu verknüpfen. Ontologien bestehen also aus dem Vokabular, mit dem Objekte und ihre Eigenschaften und Beziehungen festgehalten werden.

Ontologien können in diversen Sprachen und Formaten entworfen bzw. abgelegt werden. Die Ausdruckskraft dieser verschiedenen Ontologiesprachen wurden in zahlreichen Studien untersucht.[181] Mit der Entwicklung der Ontologiesprache OWL (s. Kapitel 6.3) durch das World Wide Web Consortium (W3C)[182] scheint diesbezüglich allerdings ein Standard erreicht zu werden. OWL-Ontologien können bereits von allen nachfolgend vorgestellten Programmen verarbeitet werden.

Es existieren zahlreiche Bearbeitungstools für Ontologien, von denen nachstehend drei zur Erläuterung einiger Arbeitsschritte im Umgang mit Ontologien vorgestellt werden.

177 Auf philosophische Definitionen des Begriffes, etwa als "Wissenschaft des Seienden" nach Aristoteles, Metaphysik IV, wird hier nicht näher eingegangen.
178 http://suo.ieee.org
179 http://ontology.teknowledge.com
180 Die englischen Begriffe hierfür sind u.a. frames/concepts, instances/objects/atoms und slots/attribute/properties.
181 [Corcho2000a], [McEntire2000]
182 http://w3.org, http://w3c.org

Kaon[183], ein an dem Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB) der Universität Karlsruhe[184] entwickeltes OpenSource Tool, ermöglicht das Importieren, Exportieren[185], Erstellen, Bearbeiten, Visualisieren und Durchsuchen von Ontologien. Abbildung 1 zeigt eine Visualisierung der BibTeX-Ontologie[186] und die Suche nach Klassen, Instanzen und Relationen.

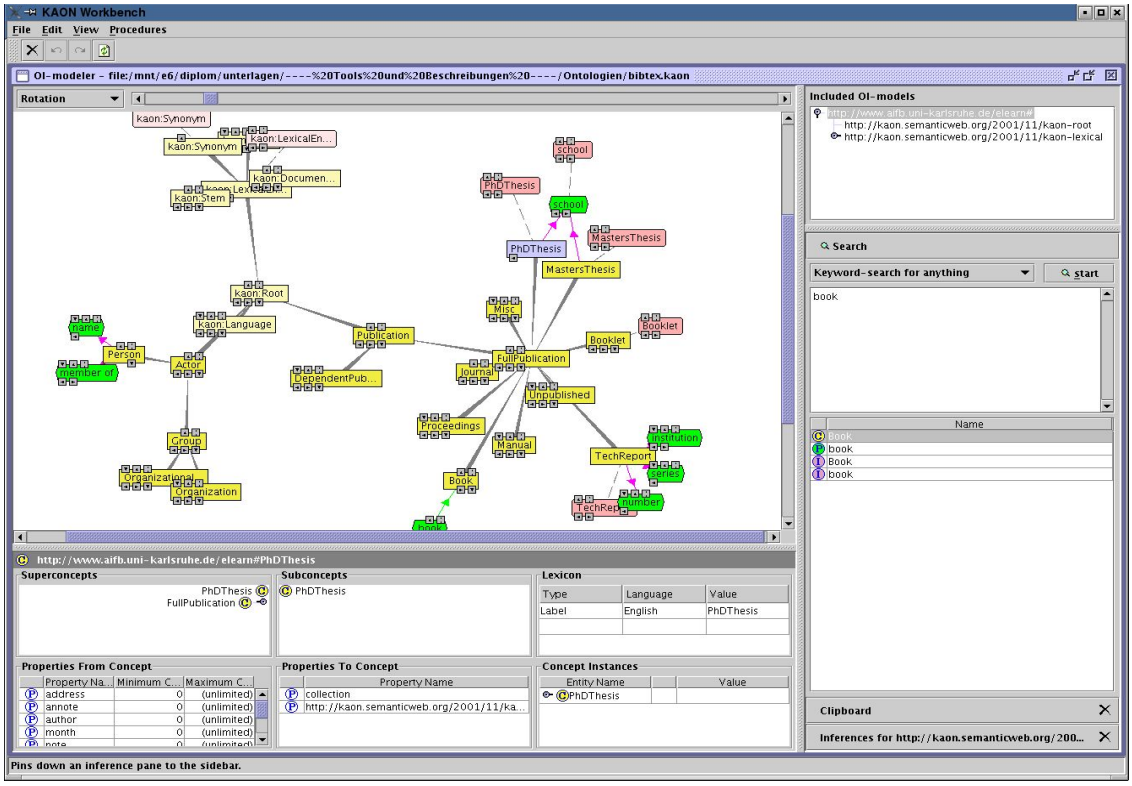

Abb. 1: KAON-Beispiel "BibTeX Ontologie"

Für das Verknüpfen von Suchargumenten besitzt KAON eine eigene Abfragesprache. Die Möglichkeit, Suchanfragen abzuspeichern, fehlt bislang jedoch, was dessen Verwendung als Suchinstrument einschränkt.

183 KAON: The KArlsruhe ONtology and Semantic Web tool suite http://kaon.semanticweb.org, http://sourceforge.net/projects/kaon/
184 http://www.aifb.uni-karlsruhe.de
185 Import: derzeit RDF(S), Protégé und KAON, Export: RDF(S), KAON.
186 "It [BibTeX] is probably the most common format for bibliographies on the Internet.", Quelle: http://www.ecst.csuchico.edu/~jacobsd/bib/formats/bibtex.html

Mithilfe des auf KAON aufbauenden "TextToOnto"-Moduls[187] können aus verschiedenen Dokumenten Terme extrahiert werden, die für das Modellieren der Ontologie infrage kommen:

> "Supports semi-automatic creation of ontologies by applying text mining algorithms. Currently includes term extraction algorithm, concept association extraction algorithm and ontology pruning algorithm. This tool aids users in creating and maintaining ontologies through application of text-mining algorithms. In such way it helps detecting concepts and relationships that the ontology engineer has initially missed, but that may be inferred from texts about the domain being modeled."[188]

Mithilfe des Ontologietools **Protégé**[189] ist es möglich, verschiedene Suchargumente zu verknüpfen und die Suche in einer Ontologie unter einem eigenen Namen abzuspeichern (s. Abb.2). Genau wie bei KAON beziehen sich bei dem an der Stanford Universität entwickelten Editor die Anfragen auf Klassen, Eigenschaften und Instanzen.

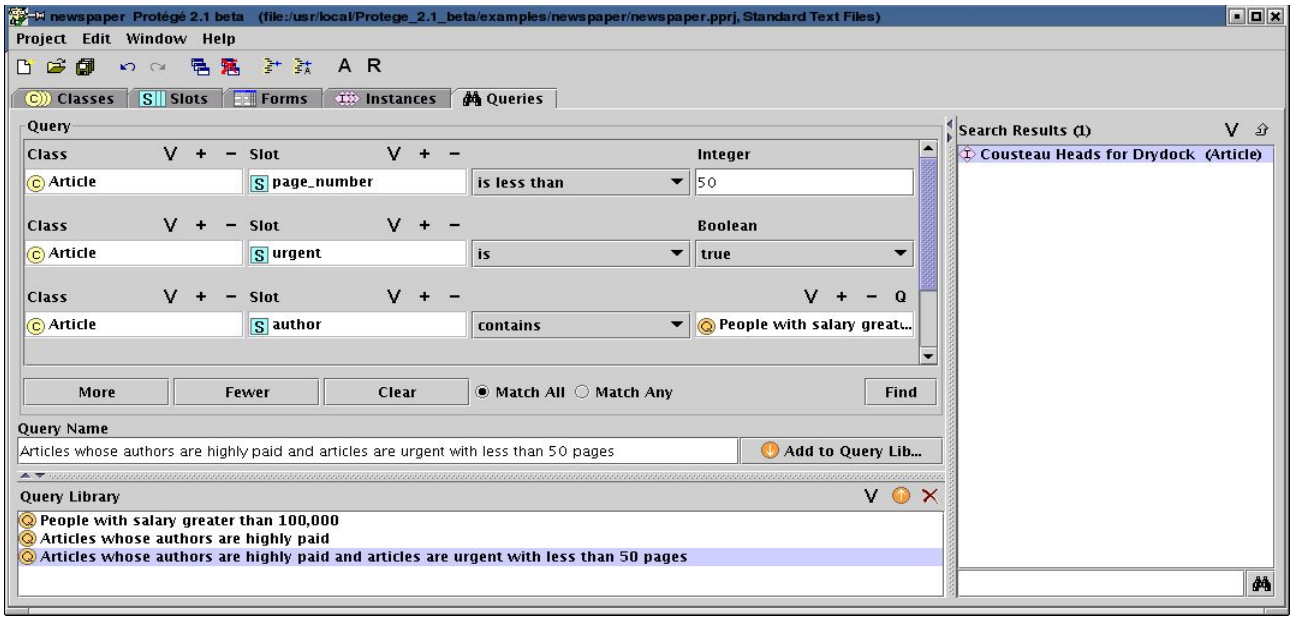

Abb. 2: Protégé-Beispiel "Suche"

Protégé bietet im Vergleich zu KAON dieselben Bearbeitungsschritte, beschränkt sich bei der Darstellung der Ontologien allerdings auf die hierarchischen Strukturen.

187 http://sourceforge.net/projects/texttoonto/
188 Quelle: http://km.aifb.uni-karlsruhe.de/kaon2/Members/rvo/Module.2002-08-22.4934
189 http://protege.stanford.edu, [Noy2000]

Mit dem ebenfalls aus Stanford stammenden **Chimaera** System[190] können Ontologien auch über das Internet bzw. Browser entworfen, bearbeitet und gepflegt werden. Zwei weitere Hauptfunktionen von Chimaera sind das Zusammenführen (merging) und die Analyse (diagnosing) von bestehenden Ontologien:

> "It supports users in such tasks as loading knowledge bases in differing formats, reorganizing taxonomies, resolving name conflicts, browsing ontologies, editing terms, etc."[191]

Abbildung 3 zeigt einige der möglichen Operationen im Hinblick auf die Bearbeitung einer hierarchischen Klassenstruktur.

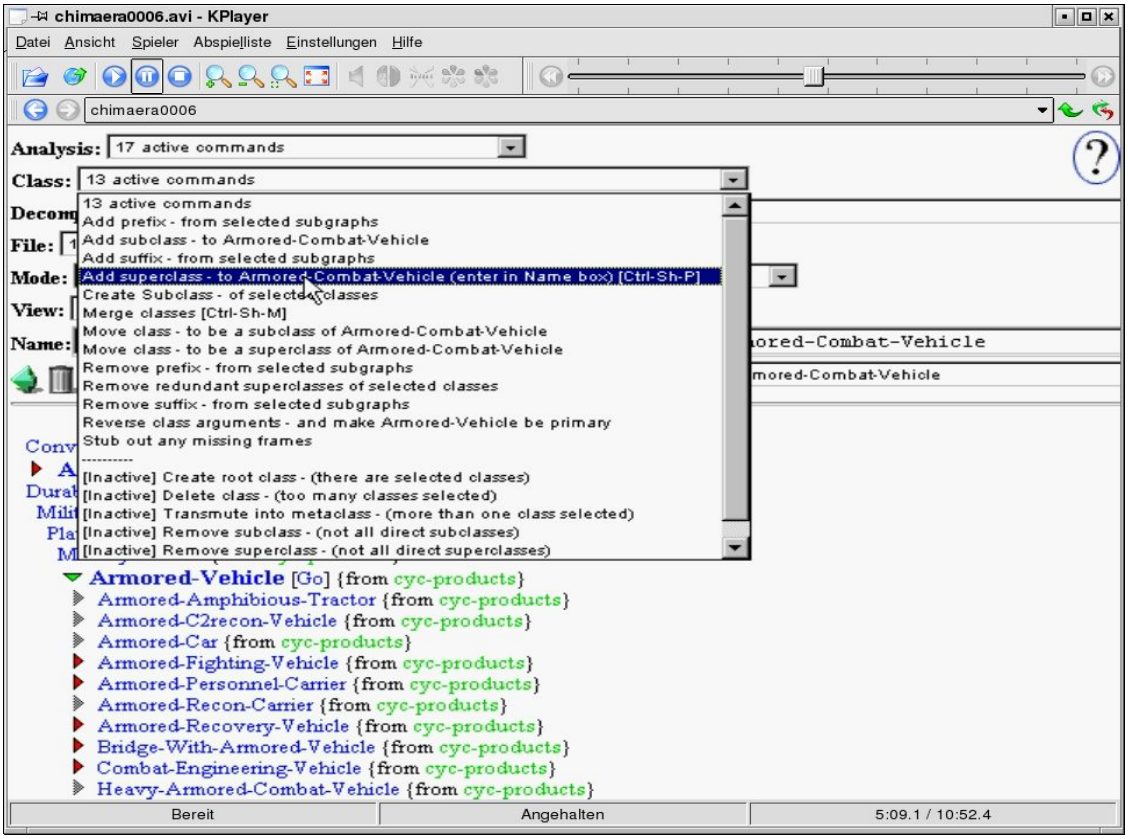

Abb. 3: Chimaera-Beispiel "Superklasse hinzufügen"

190 http://www.ksl.stanford.edu/software/chimaera/
191 Quelle: http://www-ksl-svc.stanford.edu:5915/&service=CHIMAERA

Mit allen o.g. Bearbeitungstools lassen sich auch nicht-hierarchische Relationen definieren und bearbeiten. Beispiele hierfür werden in Kapitel 6.3 gegeben.

Mögliche Anwendungsfelder von Ontologien sind der Aufbau einer Wissensbasis in Unternehmen[192] oder die Produktkatalogverwaltung.[193] Durch den Einsatz von Ontologien würden laut des W3C eine Reihe von weiteren Anwendungen profitieren:

> "Some ontology tools can perform automated reasoning using the ontologies, and thus provide advanced services to intelligent applications such as: conceptual/semantic search and retrieval, software agents, decision support, speech and natural language understanding, knowledge management, intelligent databases, and electronic commerce."[194]

Für Agentensysteme stellen Ontologien eine mögliche Hilfe bei der Darstellung und dem Vergleich von Benutzerprofilen dar.[195] Für Einkaufsagenten und Online Marktplätze können sie als Übersetzungsschicht für Anfragen dienen.[196]

192 [Kalfoglou2002]
193 [Cui2003]
194 Quelle: http://www.w3.org/TR/webont-req/#onto-def
195 [Lugo2002]
196 [Fensel2001]

6 XML-Standards

Die sog. *Extensible Markup Language* (**XML**)[197] hat sich in den letzten Jahren als generelles und vielseitiges Dokumentformat[198], sowie als Grundsprache für verschiedene Zwecke etabliert. Die derzeitige Konzentration von Softwareunternehmen auf dieses Format resultiert in einer entsprechenden Durchdringung dieses Standards. Aufgrund der vielseitigen Anwendbarkeit dieser Grundtechnologie in Bezug auf die eingangs vorgestellten Extraktions- und Darstellungsmethoden wird hier näher auf diesen Standard eingegangen.

Während XML selbst die Möglichkeit der syntaktischen Strukturierung bietet, ermöglichen auf XML aufbauende Standards die semantische Beschreibung von XML-Inhalten. Zu den XML-Kerntechologien gehören:[199]
- XSchema, DTDs, (Strukturierung der Inhalte des XML-Dokumentes)
- XPath, XPointer, XQuery (Durchsuchen der XML-Dokumente)
- XLink, XInclude (Verknüpfen, Einfügen von Dokumenten)
- XSLT (Transformierung von XML-Dokumenten)
- XSL, CSS (Darstellungsoptimierung)

Mithilfe von sog. *Namensräumen* (namespaces) lässt sich die im Dokument zu gebrauchende Taxonomie der verwendeten Tag-Namen deklarieren. Das Homonymieproblem kann dabei durch Verwendung des jeweils passenden Namensraumes geklärt werden. Die Einbindung von Namensräumen hat aber nur deklarativen Charakter, d.h. die verwendeten Tags können nicht wie bei einer DTD auf Konsistenz kontrolliert werden. Semantische Beziehungen der Tagnamen bzw. Inhalte werden damit ebenfalls nicht abgedeckt. Es existieren aber domänenspezifische Implementationen von XML, wie z.B. MathML oder CML (chemical markup language) und auf XML aufbauende Standards, welche die Beschreibungsmöglichkeiten von XML entsprechend erweitern.

Im Folgenden wird zunächst die Struktur von XML-Dokumenten erklärt, auf dessen Basis die Wissensobjekte dargestellt und gesucht werden können. Anschließend werden einige Ansätze des Retrievals in XML-Dokumenten erläutert. Diese nutzen teilweise die XML-Kerntechnologien und erweitern sie um TM-Vorgehensweisen. Mit ihnen lassen sich alle in XML abgebildete Dokumente, also z.B. auch in XML dargestellte Klassifikationssysteme oder Thesauri durchsuchen. Anschließend werden einige auf XML aufbauende Ontologiestandards vorgestellt. Sie stellen spezielle Verwendungsweisen der XML-Elemente (*XML-Tags*) dar.

197 http://www.w3.org/XML/
198 [Heikkinen2000]
199 [Ray2001]

6.1 Struktur

XML eröffnet verschiedene Möglichkeiten, ein Dokument zu strukturieren. Es lassen sich hierarchische sowie logisch abhängige bzw. verschachtelte Strukturen erstellen, z.B. in folgender Form:

```
<!DOCTYPE stefkoch PUBLIC "-//stefkoch//DTD diplom //DE" "diplom.dtd">
<diplomarbeit sprache="DE">
   <titel>Text Mining</titel>
   <autor> <nachname>Koch</nachname> </autor>
   <kapitel nr="1"> <seite>...</seite> </kapitel>
</diplomarbeit>
```

Die Zeichenfolge "diplomarbeit" ist dabei ein sog. *Tagnamen*, "nr" eine *Property*[200], während "DE", "1" und "Koch" *Werte* darstellen. Die Verwendung dieser Angaben kann durch eine sog. *Document Type Definition* (**DTD**)[201] vorgegeben werden. Die DTD definiert, aus welchen Teilen das XML-Dokument bestehen darf oder muss. Sie wird in einem separaten DTD-Dokument definiert. An ihr kann die Gültigkeit der Struktur des XML-Dokumentes gemessen werden. Stimmt das XML-Dokument mit den Vorgaben überein, so ist es "validiert", d.h. es entspricht der definierten Struktur.

Die sog. *XML Schema Definition* oder "XSchema" (**XSD**)[202] ist ein neuerer Ansatz zum Strukturieren von XML-Dokumenten und wird die DTDs voraussichtlich ablösen. Im Vergleich zu DTDs sind XSDs noch nicht weit verbreitet, bieten aber u.a. folgende Vorteile:

- Datentypdefinitionen sind erweiterbar und vererbbar.[203]
- Datentypen können genauer vorgegeben werden.[204]
- Eigene Datentypen können definiert werden.[205]
- Die Kardinalität von Elementen kann angegeben werden.[206]
- Namensräume werden unterstützt.

Eine durch DTDs oder XSDs definierte Struktur eines XML-Dokuments stellt bereits formelles Metawissen über den Text dar, da sie bereits Aussagen über die Inhaltstypen und über Beziehungen der Textkörper zulässt. Sie ist eine Art Grammatik der Wissenselemente, die in dem Dokument vorkommen.

---

200 Es existieren eine Reihe an anderen Bezeichnungen hierfür, u.a. "Attribut" oder "Eigenschaft".
201 http://www.w3.org/TR/REC-xml/
202 http://www.w3.org/XML/Schema
203 Neben dem Attribut "closed" können auch "open" und "refinable" für Tags vergeben werden. Sie müssen damit nicht, wie bei DTDs, komplett neu geschrieben werden.
204 Statt nur #PCDATA kann auch integer, string oder date vergeben werden.
205 Damit ist XSchema vielen Datenbanksystemen einen Schritt voraus.
206 Statt nur "genau 1mal", "mind. 1Mal", "1mal oder keinmal" und "beliebig oft" können auch genaue sowie maximale und minimale Vorkommnisse der Elementen angegeben werden.

6.2 Extraktion

Auf einzelne Elemente in der hierarchischen XML-Struktur lässt sich gezielt mit **XPath**[207] zugreifen. XPath wurde von der Sprache **XQL**[208] abgeleitet und bietet eine Syntax, mit der ein Tag oder eine Property sowie nach deren Werten gesucht werden kann. Der Suchstring "buch[@sprache = 'de']" liefert z.B. das Buch-Tag, dessen Property "sprache" der Wert "de" zugewiesen wurde. Ist "sprache" nicht als Eigenschaft, sondern ein eigener Tag, lautet die Suche "buch[sprache] = 'de'". Mithilfe von Wildcards, Booleschen Operatoren und relativen sowie absoluten Pfadangaben kann die Suche verfeinert werden. Der Ausdruck "//kapitel[.//titel and .//seite]" beschreibt z.B. die Suche nach einem Kapitel, unter dessen Nachfolgerelementen sich mindestens eine Überschrift und eine Seite befindet. Mithilfe von "=", "<" und ">" können numerische Werte verglichen werden. Numerische Operationen wie Addition, Ab-/ Aufrunden, sowie die Ermittlung von Minimal- bzw. Maximalwerten sind ebenfalls möglich.[209] Eine der Limitierungen von XPath ist, dass man nur exakte Treffer erhält, aber keine Näherungssuche wie "enthält 'anfangs'" möglich ist. Erweiterte Trunkierungs- und Kontextoperatoren fehlen ebenfalls.

Um den Funktionsumfang traditioneller Datenbankabfragesprachen wie z.B. SQL und Messenger[210] für das Durchsuchen von XML-Dokumenten verfügbar zu machen und somit ein sog. *XML Datenbank Management System* (XML-DBMS) umzusetzen, haben sich einige Ansätze gebildet. Ein Beispiel hierfür ist **QUILT**:

> "Our strategy in designing the language has been to borrow features from several other languages that seem to have strengths in specific areas. From XPath and XQL we take a syntax for navigating in hierarchical documents. From XML-QL[211] we take the notion of binding variables and then using the bound variables to create new structures. From SQL we take the idea of a series of clauses based on keywords that provide a pattern for restructuring data (the SELECT-FROM-WHERE pattern in SQL). From OQL we take the notion of a functional language composed of several different kinds of expressions that can be nested with full generality. We have also been influenced by reading about other XML query languages such as Lorel[212] and YATL."[213]

207 http://www.w3.org/TR/xpath/
208 http://www.ibiblio.org/xql/
209 Die entsprechenden Funktionen lauten "sum(node-set)", "round(number)", "floor(number)" und "ceiling(number)".
210 http://www.stn-international.de/training_center/messenger/commands/Contents.htm
211 http://www.w3.org/TR/NOTE-xml-ql/, [Deutsch1998]
212 http://www-db.stanford.edu/lore/, [Abiteboul1997]
213 [Chamberlin2000]

Auf der Basis von Quilt wird derzeit vom W3C der Standard **XQuery**[214] entwickelt, für den bereits eine Reihe an Anwendungen existieren.[215] Mit XQuery ist es u.a. möglich, Suchergebnisse zu aggregieren (durch sum, count, min, max, avg) und Dubletten zu eliminieren. Im Folgenden werden drei auf XQuery aufbauende Ansätze vorgestellt, die auf unterschiedliche Weise dessen Information Retrieval Funktionen erweitern.

> "As the only feature supporting information retrieval in XML, XQuery supports querying for single words in texts. There is no possibility for weighting or ranking, no support for vague query conditions, and no operator for relevance-oriented search. **XIRQL** fills this gap for a subset of the XQuery language."[216]

XIRQL führt folgende retrievaltechnische Erweiterungen ein:[217]

- Stringsuche
  Die Suche nach dem Inhalt eines Tags kann ungenau sein, d.h. es kann auch nach Teilen gesucht werden. Die Suche nach "//kapitel[.//titel cw[218] 'Mining']" findet z.B. Kapitel, die "<titel>Data Mining</titel>" oder "<titel>Text Mining</titel>"enthalten.
- Vagheit
  Auf der Basis von Thesauri und Klassifikationssystemen kann die Suche mit Ähnlichkeitsoperatoren wie "near", "broader", "narrower" oder "related" verfeinert werden.
- Relevanz
  Die Elemente können nach ihrer Position in der Hierarchie sortiert oder unabhängig von ihrer Verschachtelung zu Indexeinheiten zusammengefasst werden, für die auf Basis von TF-IDF Gewichtungen vergeben werden können. Einzelne Suchterme können ebenfalls gewichtet werden.

Implementiert wird XIRQL bspw. in der sog. "Hyper-media Retrieval Engine for XML" (Hyrex)[219].

214 http://www.w3.org/TR/xquery/
215 http://www.w3.org/XML/Query.html#products
216 [Fuhr2004]
217 [Fuhr2001]
218 "cw" steht für "contains word".
219 http://www.is.informatik.uni-duisburg.de/projects/hyrex/, [Fuhr2002], [Fuhr2003]

Die sog. *Flexible XML Search Language* (**XXL**)[220] verwendet Indizes, um den Inhalt der XML-Dokumente besser extrahieren zu können. In dem sog. *element path index* (EPI) werden die Positionen der Tagnamen abgespeichert, um den Zugriff zu ermöglichen. Über die in den XML-Dokumenten vorkommenden Elementnamen wird mithilfe von Stemming und TF-IDF ein sog. *element content index* (ECI) aufgebaut, der alle möglichen XML-Elemente auflistet. Zusätzlich kann bei XXL noch externes Wissen aus Ontologien herangezogen werden. Aus ihnen wird der *ontology index* (OI) erstellt,[221] der bei der semantischen Umfeldsuche zum Auffinden ähnlicher Inhalte bzw. XML-Tagnamen dient. Mithilfe dieser drei Indizes lassen sich also XML-Dokumente durchsuchen und gleichzeitig relevance ranking sowie das in Ontologien enthaltene Metawissen nutzen:

> "XXL incorporates special operators, like similar, into the query language and uses ontological information to automatically calculate the scores. [...] The similarity scores of different conditions are composed using simple probabilistic reasoning. [...] The first step of the computation determines similar terms (with relevance score $\pi 1$) to the given term based on the ontology. The second step computes the tf*idf-based relevance ($\pi 2$) of each term for a given element content. [...] The element content under consideration then satisfies the search condition with relevance ($\pi 1 \cdot \pi 2$)."[222]

220 [Theobald2002a]
221 [Theobald2003]
222 [Theobald2002b]

Auch der sog. *Text in XML*-Ansatz (**TIX**) erweitert XQuery-Abfragen unter Zuhilfenahme eines invertierten Index und Methoden des Relevance Ranking sowie der Phrasensuche:

> "XXL and XIRQL are two query languages supporting ranked queries on XML data. XXL incorporates special operators, like similar, into the query language and uses ontological information to automatically calculate the scores. Our system, on the other hand, enables the user to specify scoring function by providing them with language extensions with which user-defined functions can be plugged. XIRQL is the first to address the result duplication problem in the IR-style structured query where element type is not specified. They choose to return only those nodes of predetermined types. We decide not to impose such a limitation and present both a default way and a user-defined way of picking elements to be returned through a stack-based algorithm.[...]The major advances in TIX include (i) the ability to manage relevance scores, including score generation, manipulation, and use; and (ii) facilities for management of result granularity (necessitated because relevance may be associated with nested elements at multiple granularities)."[223]

Neben der Möglichkeit, das Relevance Ranking nach eigenen Vorstellungen abändern zu können, kann man mit TIX struktur- bzw. typunabhängig suchen, d.h. Tagnamen, Attributnamen oder deren Werte können als gleich betrachtet werden. TIX erweitert damit die Möglichkeit des XIRQL-Ansatzes, der bereits die Suche in Tag- und Attributnamen mithilfe eines Operators verbindet.[224]

Zusammenfassend optimieren die vorgestellten Ansätze sowohl die Suche in eindeutig als auch in unterschiedlich strukturierten XML-Dokumenten. Während für eindeutig strukturierte Dokumente die zielgenaue Extraktion von Elementen, Eigenschaften und Werten möglich ist, kann man bei unterschiedlichen Dokumenten die Suche durch die semantische Umfeldsuche erweitern. Thesauri, Klassifikationssysteme und Ontologien können die Suchmethoden dabei unterstützen.

223 [Al-Khalifa2003]

224 Mit dem Suchstring "~name" lassen sich bei XIRQL Tag- und Attributnamen und mit "~name = 'Inhalt'" deren Inhalte durchsuchen.

6.3 Abbildung

XML eignet sich neben der strukturierten Darstellung und Formatierung von Textdokumenten auch zur Abbildung der in Kapitel 5 vorgestellten semantischen Netze. Für die Darstellung von Clustern scheint XML derzeit kaum verwendet zu werden, die mögliche Anwendung in Form des auf XML aufbauenden sog. *Scalable Vector Graphics* (**SVG**)-Standards[225] ist jedoch möglich, wie die sog. "CLUTO"-Software demonstriert.[226]

Für die Beschreibung von Metadaten empfiehlt das W3C das sog. *Ressource Description Framework* (**RDF**)[227], das ebenfalls auf XML aufbaut. Durch *RDFSchema* (**RDFS**[228]) kann die Verwendung von RDF-Tags und deren Beziehungen zueinander festgelegt werden.

> "It [RDFS] is based on some ideas from knowledge representation (semantic nets, frames and predicate logic), but it is much simpler to implement (and also less expressive) than full predicate calculus languages such as CycL and KIF. Core classes are class, resource and property; hierarchies and type constraints can be defined (core properties are type, subclassOf, subPropertyOf, seeAlso and isDefinedBy). Some core constraints are also defined. A conclusion is that an ontology defined in RDF(S) will lack from functions and axioms, but concepts, relations and instances (as well as claims) can be easily defined."[229]

Die Abbildung eines Klassifikationssystems ist zwar, wie D. Vizine-Goetz mit einem Ausschnitt der DDC zeigt[230], auch direkt in XML abbildbar, die vordefinierten bzw. standardisierten Relationen von RDFS[231] ermöglichen allerdings eine eindeutige Verwendungsweise und machen die Erstellung einer (proprietären) DTD überflüssig. Mithilfe der sog. *RDF Query Language* (**RQL**) lässt sich desweiteren gezielt nach diesen Relationstypen suchen.[232]

225 http://www.w3.org/Graphics/SVG/
226 http://www-users.cs.umn.edu/~karypis/cluto/
227 http://www.w3.org/RDF/
228 http://www.w3.org/TR/rdf-schema/
229 [Corcho2000b]
230 http://www.loc.gov/catdir/bibcontrol/vizinegoetz_paper.html
231 Klassen und Properties können durch "rdfs:subClassOf" und "rdfs:subPropertyOf" verschachtelt werden. Instanzen werden den Klassen durch "rdf:type" zugeordnet.
232 [Karvounarakis2002]

Folgender Ausschnitt aus der Produkt- und Dienstleistungsklassifikation von Dun&Bradstreet stellt ein Beispiel für die Verwendung der Hierarchierelationen dar:

> "<rdfs:Class rdf:ID="Online-database-information-retrieval-systems">
> <rdfs:subClassOf rdf:resource="#Information-centers"/>
> <unspsc-code> 83121604</unspsc-code>
> </rdfs:Class>"[233]

Der sog. *XML Topic Map*-Standard (**XTM**)[234] stellt eine einfache DTD für XML-Dokumente dar. XTM bietet nur wenige Strukturrelationen, die durch eigene Tags definiert werden. Die RDF-Relation "rdf:type" wird so bspw. durch einen eigenen "<instanceOf>"-Tag abgebildet.

Für die Darstellung von Thesauri in XML existieren derzeit unterschiedliche Formate[235]. Das W3C arbeitet derzeit allerdings an einer diesbezüglich einheitlichen Darstellung mithilfe von RDFS, dem sog. S*imple Knowledge Organisation System* (SKOS).[236]

Aufgrund der vielfältigen Möglichkeiten von Ontologien zur Strukturierung von Texten werden nachfolgend einige Ontologiestandards erläutert. Webbasierte Ontologiesprachen versuchen, Webstandards wie XML und RDF mit traditionellen Ontologiesprachen zu vereinigen. Sie haben dabei das Ziel, die Beschränkungen von RDF und XML bezüglich der Darstellung semantischer Beziehungen aufzuheben:

> "Although XML DTDs and XML Schemas are sufficient for exchanging data between parties who have agreed to definitions beforehand, their lack of semantics prevent machines from reliably performing this task given new XML vocabularies. The same term may be used with (sometimes subtle) different meaning in different contexts, and different terms may be used for items that have the same meaning. RDF and RDF Schema begin to approach this problem by allowing simple semantics to be associated with identifiers. With RDF Schema, one can define classes that may have multiple subclasses and super classes, and can define properties, which may have sub properties, domains, and ranges. In this sense, RDF Schema is a simple ontology language. However, in order to achieve interoperation between numerous, autonomously developed and managed schemas, richer semantics are needed. For example, RDF Schema cannot specify that the Person and Car classes are disjoint, or that a string quartet has exactly four musicians as members."[237]

233 Quelle: http://www.ksl.stanford.edu/projects/DAML/UNSPSC.daml
234 http://topicmaps.org/xtm/
235 http://www.w3c.rl.ac.uk/SWAD/thes_links.htm
236 http://www.w3.org/2001/sw/Europe/reports/thes/
237 Quelle: http://www.w3.org/TR/webont-req/

Die webbasierten Ontologiesprachen besitzen im Vergleich zu reinen Ontologiesprachen nicht dieselben Modellierungs- und Aggregationsfähigkeiten,[238] erweitern aber die Darstellungsmöglichkeiten der herkömmlichen Webstandards wie XMLSchema und RDFSchema, indem sie bspw. vorgefertigte Definitions-, Komplementär- sowie Disjunktions- und Konjunktionsrelationen zur Verfügung stellen.[239]

Ein früher Ansatz zur Benutzung von Ontologien im WWW waren die sog. *Simple HTML Ontology Extensions* (**SHOE**)[240], durch die HTML-Dateien mit Kategorien und Definitionen erweitert und gesucht werden können. Die Entwicklung von SHOE wurde jedoch zugunsten der folgenden Standards beendet. Die *Ontology Modelling Language* (**OML**)[241] stellt eine auf XML angewandte Form von SHOE dar.

Um noch weitere Relationen darstellen zu können, hat die DARPA die sog. *Darpa Agent Markup Language* (**DAML**)[242] auf der Basis von RDF entworfen. Die meisten der o.g. Tools können DAML-Ontologien importieren und exportieren.

Die sog. *Ontology Inference Layer* (**OIL**)[243], hervorgegangen aus dem On-To-Knowledge-Projekt[244] ist als Folgerungsschicht von Ontologien konzipiert worden und kombiniert Modellierungsmethoden rahmenbasierter (framebased) Sprachen mit formaler Semantik[245] und Beschreibungslogik (description logics). OIL verwendet für die formale Semantik und Syntaxdefinitionen ebenfalls RDF und basiert auf Ideen von XOL und OKBC.[246]

OIL und DAML werden mittlerweile zusammengefasst und als **DAML+OIL**[247] bezeichnet.

238 [Su2002]
239 [Gil2000]
240 http://www.cs.umd.edu/projects/plus/SHOE/
241 http://www.ontologos.org/OML/OML 0.3.htm
242 http://www.daml.org
243 http://www.ontoknowledge.org/oil/
244 http://www.ontoknowledge.org
245 [Horrocks2000]
246 [Fensel2000]
247 http://www.w3.org/TR/daml+oil-reference

Den Nachfolger von DAML+OIL stellt die sog. *Ontology Web Language* (**OWL**)[248] dar. Sie ist ebenfalls eine Vokabularerweiterung für RDF und ist seit Februar 2004 Empfehlung des W3C.

Der volle Sprachumfang von OWL ("OWL Full") kann auf Untermengen begrenzt werden, um Ansprüchen der Aussagenlogik zu genügen ("OWL Description Logics") oder durch Reduzierung der Komplexität eine einfachere Implementierung in Software zu ermöglichen ("OWL Lite"). Im Anhang wird ein Beispiel für die unterschiedlichen Schreibweisen von DAML und OWL gegeben. OWL erweitert o.g. RDF-Relationen, indem z.B. Restriktionen in Bezug auf Kardinalität einer Klasse definiert werden können.[249] Desweiteren lassen sich Klassen, Eigenschaften, Instanzen und Werte näher beschreiben.[250] Im Gegensatz zu DAML+OIL können mit OWL z.B. auch symmetrische Relationen definiert werden. Die Synonymierelation wird in Bezug auf Instanzen mithilfe von "sameAs" und in Bezug auf Klassen mithilfe von "equivalentClass" beschrieben:

```
<owl:Description rdf:about="#William_Jefferson_Clinton">
   <owl:sameAs rdf:resource="#Bill_Clinton"/>
</owl:Description>
```

Die Antonymierelation kann durch "differentFrom" ausgedrückt werden:

```
<owl:Class rdf:about ="person">
</owl:Class>
<person rdf:ID="Bill_Clinton">
   <owl:differentFrom rdf:resource="#George_Bush"/>
</person>
```

248 http://www.w3.org/TR/owl-ref/

249 Durch Verwendung der Properties "maxCardinality", "minCardinality", "cardinality" muss für diesen Zweck nicht auf eine XSD zurückgegriffen werden.

250 Die Rolle der Klassen können näher durch "subClassOf", "equivalentClass", "complementOf", "unionOf", "intersectionOf" beschrieben werden.
Properties lassen sich durch "subPropertyOf", "equivalentProperty", "disjointWith", "inverseOf", "FunctionalProperty", "InverseFunctionalProperty", "TransitiveProperty" sowie "SymmetricProperty" spezifizieren.
Für Instanzen können die Relationen "sameAs", "differentFrom" und "AllDifferent" definiert werden.

Neben den vordefinierten Relationen können auch eigene Relationen bzw. Properties definiert werden. So kann bspw. definiert werden, dass die einem Objekt zugeordnete Eigenschaft "gewähltVon" inversiv zu der Relation "wählt" ist:

```
<owl:ObjectProperty rdf:ID="wählt">
<owl:ObjectProperty rdf:ID="gewähltVon">
   <owl:inverseOf rdf:resource="#wählt" />
</owl:ObjectProperty>
```

Desweiteren kann z.B. angegeben werden, dass nur Frauen "First Ladies" von Präsidenten sein können:

```
<owl:ObjectProperty rdf:ID="firstLady">
   <rdf:type rdf:resource="&owl;FunctionalProperty" />
   <rdfs:domain rdf:resource="#präsident" />
   <rdfs:range rdf:resource="#frau" />
</owl:ObjectProperty>
```

Ein weiteres Beispiel zeigt, dass auf die Eigenschaften verschiedene Restriktionen gelegt werden können. So kann z.B. definiert werden, dass die Relation "hatFirstLady" sich nur auf Präsidenten bezieht und ein Präsident nur mit genau einer Frau diese Beziehung haben kann:

```
<owl:Class rdf:ID="präsident">
   <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource="#hatFirstLady" />
         <owl:allValuesFrom rdf:resource="#präsident" />
      </owl:Restriction>
   </rdfs:subClassOf>
   <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource="#hatFirstLady" />
         <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger"> 1
         </owl:cardinality>
      </owl:Restriction>
   </rdfs:subClassOf>
</owl:Class>
```

7 Transfer

Damit das dargestellte Wissen analysiert und nutzbar gemacht werden kann, muss der Zugriff auf das Wissen und dessen Abgleich ermöglicht werden. Erst dadurch wird Knowledge Discovery in Text möglich. Nach F. Rötzer stellt der Wissensaustausch die Grundlage der Wissensgesellschaft dar.[251]

Damit das Wissen wiederverwendbar wird, muss es auch auf anderen Plattformen ohne viel Aufwand integriert werden können, d.h. ohne dass dadurch bestehende Abläufe gestört werden. Hierfür sind Schnittstellen nötig, welche die zur Wissensrepräsentation verwendeten Systeme verbinden. Als Schnittstelle für den Zugriff auf Knowledge Representation Systeme[252] verschiedener Typen entwickelte die Stanford University das *Open Knowledge Base Connectivity* (**OKBC**) Protokoll[253], das auch in einigen der Ontologietools als Austauschschicht dient.[254] Für die Kommunikation von Agentensystemen wird die bereits erwähnte Knowledge Query and Manipulation Language (KQML) verwendet. Hier können Ontologien bei der Übersetzung der Nachrichten eine wichtige Rolle spielen.[255] Das sog. *Knowledge Interchange Format* (**KIF**)[256], auf dessen Basis auch die Suggested Upper Merged Ontology entwickelt wurde, stellt dabei einen grundlegenden Standard für den Wissensaustausch dar, wird hier aber zugunsten von Standards zum Wissenstransfer über das Internet nicht weiter erläutert.

Aufbauend auf den Internet-Basisdiensten wie News, Email und WWW sollen webbasierte Content Management Systeme, WBT-Systeme oder das sog. "Semantic Web" Umgebungen für diesen Transfer schaffen.[257] Die drei nachfolgend vorgestellten Ansätze nehmen entsprechend ihrer Reihenfolge an Strukturierungsmöglichkeiten des Wissens sowie an der Mächtigkeit der definierten Schnittstellen für das in den Dokumenten dargestellte Wissen zu.

---

251 [Rötzer1999]

252 "By a KRS we mean both systems that would traditionally be considered KRSs, as well as can be viewed as a KRS, for example, an object-oriented database." in: [Chaudhri1998a]

253 [Chaudhri1998b]

254 "At SRI, OKBC bindings were defined for LOOM, Theo, SIPE-2, and Ocelot. At Stanford KSL, an OKBC server has been implemented for Ontolingua, ATP (a theorem prover), file system KB, Tuple-KB and CLOS. The University of Southern California's Information Sciences Institute has now produced its own version of an OKBC binding for LOOM. An OKBC binding for Cyc has been defined by Cycorp. The Section of Medical Informatics (SMI) at Stanford has built an OKBC server for their system Protege. Quelle: " http://www.ai.sri.com/~okbc/okbc-faq/Implementations/existing_compliant_server.htm

255 [Takeda1995]

256 http://suo.ieee.org/SUO/KIF/, http://logic.stanford.edu/kif/dpans.html

257 Die Ansätze bezüglich des Transfers von Wissen durch Content Management Systeme und WBT-Systeme können aufgrund ihrer Anzahl hier nicht umfassend diskutiert werden.

Das eingangs erwähnte Content Management System Wiki eignet sich aufgrund seiner nur rudimentär implementierten Zugriffsbeschränkungen für eine schnelle und unkomplizierte Veröffentlichung von Wissen im Intra- sowie im Internet. Durch die Offenheit des Systems kann Jeder die Beiträge verändern. Es existieren allerdings kaum Formatierungsregeln und nur wenige Strukturierungsmöglichkeiten für die Artikel. Die internen Verknüpfungen stellen das Hauptinstrument für den Zugriff auf das im System vorhandene Wissen dar.

Die vom US-Verteidigungsministerium gesponsorte *Advanced Distributed Learning* (**ADL**) Initiative[258] beschäftigt sich mit der Interoperabilität von WBT-umgebungen und -inhalten, um das Ziel des ubiquitären und personalisierbaren Zugriffs auf wiederverwendbare Inhalte zu erreichen. Zu diesem Zweck hat die Initiative das sog. *Sharable Content Object Reference Model* (**SCORM**) verabschiedet:

> "SCORM currently provides an Application Programming Interface (API) for communicating information about a learner's interaction with content objects, a defined data model for representing this information, a content packaging specification that enables interoperability of learning content, a standard set of meta-data elements that can be used to describing learning content and a set of standard sequencing rules which can be applied to the organization of the learning content."[259]

Die letzte Version stellt die SCORM Version 1.3 dar, die im Januar 2004 verabschiedet wurde. Seit der Version 1 wurde ein Schwerpunkt auf XML-basierte Metadatenkontrolle gelegt. Anwendung findet dieser Standard u.a. im sog. *Instructional Management System* (**IMS**) des Global Learning Consortium[260]. Die sog. *SCORM Conformance Test Suite*[261] ermöglicht das Überprüfen der SCORM-gerechten Verwendung von IMS Metadaten in der eigenen Lernumgebung.

258 http://www.adlnet.org
259 http://www.adlnet.org/index.cfm?fuseaction=scormabt
260 http://www.imsglobal.org, http://www.imsproject.org
261 http://www.adlnet.org/index.cfm?fuseaction=SCORDown&listing=Software

Mithilfe der im Laufe der Entwicklung zustande gekommenen sog. *Vocabulary Definition Exchange* (**VDEX**) Spezifikation lassen sich hierarchische Strukturen der verwendeten Terminologie darstellen und austauschen:

> "VDEX defines a grammar for the exchange of simple machine-readable lists of values, or terms, together with information that may aid a human being in understanding the meaning or applicability of the various terms. VDEX may be used to express valid data for use in instances of IEEE LOM [262], IMS Metadata, IMS Learner Information Package and ADL SCORM, etc, for example. In these cases, the terms are often not human language words or phrases but more abstract tokens. VDEX can also express strictly hierarchical schemes in a compact manner while allowing for more loose networks of relationship to be expressed if required."[263]

Mithilfe von VDEX kann das in den Lernumgebungen verwendete Vokabular angeglichen werden. Für die Abbildung von komplexen Zusammenhängen scheinen die hier erwähnten "loose networks" zwar nicht geeignet zu sein, für einen erfolgreichen Abgleich der Lerneinheiten stellen sie allerdings einen grundlegenden Schritt dar.

Auch das W3C sieht die Verwendung eines gemeinsamen Vokabulars als grundlegend für den Austausch von Wissen an. Der sog. **Semantic Web**-Ansatz[264] versucht, Wissensaustausch im WWW auf Basis von gemeinsamen Ontologien und eindeutig festgelegten Schlussfolgerungsregeln (Logik) zu ermöglichen. Für beide Bereiche arbeitet das W3C derzeit Empfehlungen aus, die Standards gleichkommen. Ontologien stellen im Semantic Web die Zwischenschicht dar, mit deren Hilfe die Daten in semantische Zusammenhänge gebracht werden:

> "Ontologies figure prominently in the emerging Semantic Web as a way of representing the semantics of documents and enabling the semantics to be used by web applications and intelligent agents. Ontologies can prove very useful for a community as a way of structuring and defining the meaning of the metadata terms that are currently being collected and standardized. Using ontologies, tomorrow's applications can be "intelligent," in the sense that they can more accurately work at the human conceptual level. Ontologies are critical for applications that want to search across or merge information from diverse communities."[265]

262 http://ltsc.ieee.org/wg12/
263 http://www.imsglobal.org/vdex/
264 http://www.w3.org/2001/sw/
265 http://www.w3.org/TR/webont-req/#onto-def

Gemäß Abbildung 4[266] besteht das Semantic Web grundsätzlich aus:
- einem Regelwerk für das Auffinden (URI)[267] und Kodieren der Objekte.
- einem Datenmodell (bspw. XML),
- einem Relationsmodell (bspw. RDF, RDFS oder UML[268]),
- einem semantischen Netz (bspw. DAML+OIL oder OWL),
- einem Regelwerk für Schlussfolgerungen (bspw. SWRL oder F-Logic[269]) und
- einer Prüfungsinstanz.

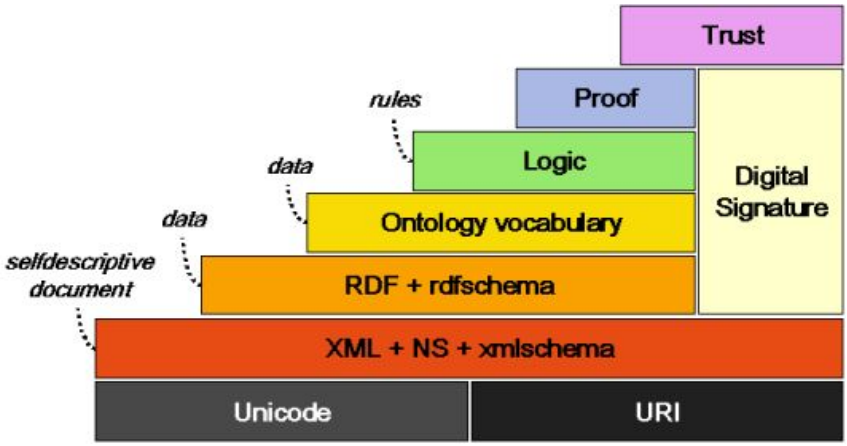

Abb. 4: Schichten des Semantic Web nach T. Berners-Lee

Aufbauend auf der über eine eindeutige Zugriffsbeschreibung auffindbaren Datenbasis bedarf es noch der Logik (**Logic**), um Aussagen aus dem Wissen extrahieren zu können. Eine Prüfungsinstanz (**Proof**) soll das entsprechende Vertrauen (**Trust**) in die produzierten Aussagen ermöglichen (s. Kapitel 9). T. Berners-Lee schlägt diesbezüglich vor, dass sich Benutzer, die Änderungen in der Wissensbasis vornehmen, durch eine digitale Signatur ausweisen.[270]

Damit die Regeln der Aussagenlogik auf das in den Ontologien abgebildete Wissen angewandt werden können, müssen die dort dargestellten Relationen als einfache Aussagen aufgefasst werden. Durch Verknüpfung der Aussagen lassen sich dann neue Aussagen herleiten.

Die bereits erwähnte "common sense reasoning engine" von Cycorp Inc. wendet Inferenzregeln auf die aus dem RKF-Projekt hervorgegangenen DAML-Ontologien und auf XTM-Taxonomien an. Ein Ansatz für die Verbindung von XML (bzw. XPath) mit den logikorientierten Sprachen Prolog und F-Logic existiert bereits mit dem sog. *XPathLog*.[271]

266 Quelle: http://www.w3.org/2001/12/semweb-fin/w3csw/
267 URI: http://www.w3.org/Addressing/, http://www.ietf.org/rfc/rfc2396.txt
268 [Sherif2002]
269 [Kifer1990]
270 http://www.w3.org/Signature/
271 [May2001]

Der Vorschlag des W3C bezüglich einer derartigen Verknüpfung stellt die sog. *Semantic Web Rule Language* (**SWRL**) dar, die sich derzeit allerdings noch in der Entwicklung befindet.[272] Sie erweitert die Fähigkeiten von OWL-Ontologien durch die Aussagenlogik von RuleML[273], sodass in den Ontologien auf einer konzeptuellen Ebene gesucht werden kann und aus den Daten mithilfe von Inferenzregeln neue Aussagen hergeleitet bzw. "bewiesen" werden können. Die in einer OWL-Ontologie festgelegten Vokabulareinträge x,y,z und deren Relationen zueinander können so z.B. wie folgt verknüpft werden:

"BruderVon(x,z) $\wedge$ SchwesterVon(y,z) $\Rightarrow$ BruderVon(x,y)".

Damit der Transfer von in zwei unterschiedlichen Ontologien festgehaltenem Wissen funktioniert, müssen diese entweder zu einer neuen zusammengeführt oder ein sog. *Mapping* zwischen ihnen erstellt werden, das als Zwischenschicht synonyme Klassen o.ä. übersetzen kann. Dadurch können beide Ontologien bestehen bleiben und müssen nicht transformiert werden. Die von der EU geförderte sog. *SDK-Cluster*-Arbeitsgruppe[274] beschäftigt sich u.a. mit den Schwierigkeiten bei dem Abgleich (mediation) von Ontologien. Für das Aufstellen von Mapping-Regeln können verschiedene Verfahren angewandt werden.[275] Der sog. *GLUE*-Ansatz[276] verwendet z.B. den Jaccard-Koeffizienten zur Ähnlichkeitsbestimmung von Konzepten. Durch Kombination mehrerer Verfahren lassen sich die Mappings verbessern.[277] Ein Tool, mit dem Mappings zwischen Ontologien erstellt werden können, ist "OntoLink"[278].

Aufbauend auf OML und Ontolingua[279] wurde die sog. X*ML Ontology Exchange Language* (**XOL**)[280] entworfen. Sie stellt eine Austauschsprache für Ontologien dar, mit deren Hilfe Ontologien abgeglichen werden können. Die Entwicklung von XOL stellt den Versuch dar, das OKBC-Wissensmodell (als OKBC-Lite) mithilfe von XML auszudrücken.

Um den Wissensaustausch in einem sog. Peer-to-Peer-Netzwerk (P2P) zu ermöglichen, verwenden die Projekte "Science to Science (S2S)"[281] und "Edutella"[282] bereits einige der o.g. Semantic Web Technologien.

---

272 http://www.daml.org/rules/proposal/
273 http://www.ruleml.org
274 http://www.sdkcluster.org, http://sdk.semanticweb.org
275 [Sure2004], S.5-6
276 [Doan2002]
277 [Sure2004], S.12-13
278 MINDSWAP: Maryland Information and Network Dynamics Lab Semantic Web Agents Project http://www.mindswap.org/2004/OntoLink/
279 http://www.ksl.stanford.edu/software/ontolingua/
280 http://www.ai.sri.com/pkarp/xol/
281 http://s2s.neofonie.de
282 http://edutella.jxta.org

8 Integration

Das Wissen muss, wenn es genutzt werden soll, nicht nur aus den Dokumenten bzw. Wissensplattformen extrahiert und analysiert, sondern auch in geeigneter Form in vorhandene Informationsinfrastrukturen integriert werden. Die Integration von Wissensquellen sollte dem Ideal eines sog. *Data Warehouse* folgen, das "[...] den kontrollierten Zugang zu benutzergerechten Daten mit der Flexibilität und Wirtschaftlichkeit der Selbstbedienung[...]"[283] ermöglicht. Für die Integration von Daten in die entsprechenden Benutzerumgebungen[284] sieht W. Martin die Funktionen Extraktion, Transformation, Aufbereitung, Verwaltung, Bündelung und Aggregation von Daten als Voraussetzung.[285]
K. Wilde weist diesbezüglich auf zwei grundlegende Problembereiche hin:

> "Die Zusammenführung der Daten [...] erfordert eine syntaktische Datenintegration. [...] treten darüberhinaus semantische Inkonsistenzen auf."[286]

Zur Lösung beider Problembereiche schlägt Wilde die Verwendung einer gemeinsamen **Metadatenbank** mit einheitlichen Formaten und normiertem Vokabular vor:

> "Man legt über die Data Warehouse Architektur eine Metadaten-Schicht, die nicht nur die Metadaten des Data Warehouse enthält, sondern auch die Metadaten von Text-Verwaltungs- und Speicherungssystemen, die in die erweiterte Data Warehouse Lösung aufgenommen werden sollen. Im Prinzip kann so jede komplexe Datentype angefügt werden. So wird aus einem Data Warehouse ein 'Knowledge Warehouse'."[287]

Cyc Inc. verspricht mit seinem sog. *Semantic Integration Bus*[288] eine Anbindung von Daten aus DBMS, WWW, Texten und Bildern in die Cyc Knowledgebase. Welche Verfahren dort zum Einsatz kommen, kann hier nicht geklärt werden. Während sich der Cyc Ansatz als Schnittstelle versteht, gibt es am Markt eine Reihe von sog. *Portallösungen*, die neben der Integration von Daten eine Benutzerschicht zur Verfügung stellt, die den Zugriff und die Bearbeitung der Quellen ermöglicht.

283 [Kirchner1998]
284 Ein "Text Warehouse" sollte bspw. Textverarbeitungs- und Emailprogamme integrieren können.
285 [Stock2000a], S.51
286 [Wilde2001], S.7f.
287 [Martin1998]
288 http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc/

Für den Aufbau einer integrativen Metadatenbank stellen Klassifikationen, Cluster, Thesauri oder Ontologien geeignete Meta-Informations-Strukturen bereit, die als Grundlage für den Index der Datenbank verwendet werden können. Klassifikationen sowie Suchmaschinen des WWW können in geeigneter Form in das Unternehmensportal integriert werden.

In Ontologien lässt sich zwar Wissen direkt darstellen, solange diese aber nicht für die Erstellung der Wissensobjekte verwendet werden und eine Überführung der Dokumente in eine Ontologie nicht möglich ist, können sie lediglich als Indexierungshilfe dienen. Angewendet bedeutet dies ein Anreichern (des Indexes) der Metadatenbank oder der Dokumente mit Ontologie-Informationen. Für Ersteres wurde bereits ein Ansatz vorgestellt. Für die sog. *Annotation* von Dokumenten, d.h. die Anreicherung mit Ontologie-Metadaten, gibt es eine Reihe von Verfahren und Anwendungen. Die Firma Teknowledge Corporation bietet bspw. Tools an, mit denen Microsoft Word und Powerpoint-Dokumente mit Ontologieeinträgen versehen werden können.[289]

Neben der Erstellung eigener Ontologien lassen sich auch vorhandene integrieren. Es existieren mittlerweile viele einschlägige Ontologien für verschiedene Anwendungsbereiche.[290] Bei der Auswahl einer Ontologie müssen vor allem dessen Darstellungsmöglichkeiten in Bezug auf das eigene (Domänen-)Wissen betrachtet werden.[291] Hierfür muss neben der thematischen Übereinstimmung mit dem geplanten Einsatzbereich geprüft werden, ob die Ontologie zu speziell bzw. zu allgemein ist und ob sie sich trotzdem integrieren lässt.[292]

Für die Integration von Ontologien in das eigene Datenbank Management System kann man z.B. das Tool "OntoSQL"[293] verwenden. Basierend auf dieser Anwendung bietet der "OntoAgent"[294] die für die Integration von Ontologien notwendigen Abgleichmechanismen. Werden einige der frei im WWW verfügbaren Ontologien[295] verwendet, so kann man mit "OntoView"[296] die Veränderungen dieser Ontologien im Auge behalten.[297]

289 MS Word: SemanticWord, MS Powerpoint: Briefing Associate http://mr.teknowledge.com/DAML
290 [Gómez-Pérez2004]
291 [Compton1996]
292 [Reategui1997]
293 http://www.aifb.uni-karlsruhe.de/WBS/aeb/ontosql/
294 http://www.i-u.de/schools/eberhart/ontoagent/, [Eberhart2002]
295 Eine Liste ist z.B. auf http://www.daml.org/ontologies/ zu finden.
296 http://ontoview.org
297 [Klein2002]

Der sog. "(ONTO)2Agent" kann dabei helfen, Ontologien im Netz ausfindig zu machen:

> "As a first step to solving the problem of searching for candidate ontologies, we present (ONTO)2Agent, an ontology-based WWW broker on the field of ontologies that spreads information about existing ontologies, helps to search appropriate ontologies, and reduces the search time for the desired ontology. (ONTO)2Agent uses as a source of its knowledge an ontology about ontologies (called Reference Ontology) that plays the role of a yellow pages of ontologies. [...] For example, when a knowledge engineer is looking for ontologies written in a given language applicable to a particular domain, (ONTO)2Agent can help in the search, supplying the engineer with a set of ontologies that totally/partially comply with the requirements identified."[298]

Das Unternehmen Ontoprise[299] deckt nach eigenen Angaben mit ihrer Produktpalette folgende Funktionen im Umgang mit Ontologien ab:

- OntoEdit[300]®: Ontologien erstellen, modifizieren und darstellen.
- OntoAnnotate®: Wissen erstellen, pflegen, teilen und verfügbar machen.
- OntoBroker®: deduktives Datenbanksystem, inferiert über F-Logik.
- SemanticMiner®: Knowledge Retrieval Plattform.
- OntoOffice: Integration von Ontologien während der Texteingabe in MS WordTM, OutlookTM oder ExcelTM.

Eine mögliche Integration in eine Portalumgebung sowie das Zusammenspiel einzelner Ontologietools und -standards zeigt das sog. "On-To-Knowledge-Tool-Repository"[301].

Soll das Retrieval in eigenen Textmengen durch Ontologien verbessert werden,[302] so sind vorhandene domänenspezifische Ontologien nur in Ausnahmefällen hilfreich. Hier bietet sich eher die Integration von thematisch allgemeinen oder linguistischen Ontologien wie Cyc oder WordNet an, um die Vielfalt der Thematiken bzw. grammatikalischen Relationen einbeziehen zu können. WordNet kann z.B. als OWL-Ontologie bezogen werden.[303] Soll das Wissen direkt in einer Ontologie hinterlegt werden, so bietet sich die Cyc-Ontologie als Grundlage an, auf der aufgesetzt werden kann.

298 [Vega1998]
299 http://www.ontoprise.de
300 [Sure2003], [Pinto2004]
301 http://www.ontoknowledge.org
302 [Aitken2000]
303 Die sog. "Knowledge, Information and Data Processing Group" bietet WordNet unter http://taurus.unine.ch/GroupHome/knowler/wordnet.html an.

Ein Beispiel für eine Plattform, auf dessen Basis Data Mining-Verfahren mit Text Mining-Algorithmen erweitert werden können, ist die frei verfügbare sog. *Waikato Environment for Knowledge Analysis* (WEKA)[304]. Für sie existieren mittlerweile eine Reihe an Text Mining Anwendungen,[305] u.a. die sog. *General Architecture for Text Engineering* (GATE)[306].

Die wohl bekanntesten kommerziellen Text Mining Anwendungen, die sich mehr mit der Analyse von Text befassen, sind der "IBM Intelligent Miner for Text"[307], der "SAS Text Miner"[308] sowie das "SPSS Text Mining for Clementine®"[309]. Alle drei benutzen laut Produktbeschreibungen ausschließlich Klassifizierung und Clustering zum Ordnen der Texte. Ob sie mittlerweile auch Taxonomien, Thesauri, Topic Maps und Ontologien als Hilfsmittel einsetzen, kann im Rahmen dieser Arbeit nicht geklärt werden.

Für den deutschsprachigen Raum etablieren sich derzeit eine Reihe an Softwareunternehmen, die ebenfalls versuchen, einige Aspekte des Text Mining abzudecken. Neben dem bereits angesprochenen Unternehmen Ontoprise, welches sich auf die Semantic Web Technologien konzentriert, sei hier auf die g.a.d.t GmbH[310] und die moresophy GmbH[311] hingewiesen. Das Spezialgebiet des erstgenannten Unternehmens stellen semantische und syntaktische Textanalysen dar, während sich letzteres auf die Anwendungsmöglichkeiten semantischer Netze konzentriert.

304 http://sourceforge.net/projects/weka/, http://www.cs.waikato.ac.nz/~ml/
305 Term Extraction: http://www.nzdl.org/Kea/
Clustering, WordNet, POS Tagging/Parsing: http://www.d.umn.edu/~tpederse/code.html
306 http://gate.ac.uk/
307 http://www.ibm.com/software/data/iminer/fortext/
308 http://www.sas.com/technologies/analytics/datamining/textminer/
309 http://www.spss.com/lexiquest/text_mining_for_clementine.htm
310 http://gadt.de
311 http://www.moresophy.de

9 Zusammenfassung und Ausblick

Aus dem breiten Themenspektrum des Text Mining wurden zunächst einige grundlegende Ansätze der Extraktion von Wissen vorgestellt (Kapitel 3). Anschließend wurden einige für das Text Mining notwendige Analyse- und Darstellungsverfahren erläutert, die ihren Ursprung in dem Feld der Knowledge Discovery in Databases und der KI haben (Kapitel 4, 5). Es wurde gezeigt, welchen Beitrag XML-basierte Verfahren und Standards zur Umsetzung der Extraktion und Darstellung von Wissen derzeit leisten können (Kapitel 6). Anschliessend wurde der Semantic Web Ansatz vorgestellt, bei dem einige dieser Standards für den Wissenstransfer Anwendung finden (Kapitel 7).
Die Arbeit schließt mit einigen Überlegungen bezüglich der Implementierung der vorgestellten Ansätze in die eigene Informationsinfrastruktur und weist auf einige Softwareprodukte hin, die dies ermöglichen (Kapitel 8).

Damit wurde das Gebiet der Wissensextraktion aus Texten grob umrissen und gezeigt, welche vielfältigen Lösungsansätze das Text Mining hierfür bietet. Die Vielfalt an vorhandenen Analyse- und Darstellungsverfahren scheint einer der Gründe dafür zu sein, dass sie bislang nur ansatzweise in Softwareprodukten implemetiert sind. Der Bedarf an den Technologien bzw. deren mögliche Einsatzbereiche werden allerdings mittlerweile erkannt. Neben den bereits erwähnten Managementinformationssystemen stellen die Bereiche Kundenbetreuung, Marketing (hier insbesonders Marktforschung und Medienwirkungsanalyse) und Workflowmanagement denkbare Anwendungs-szenarien für Text Mining- bzw. Knowledge Discovery-Methoden dar.

Die Anwendung klassischer (statistischer) Data Mining-Algorithmen im Text Mining sowie die Standardisierung der vorgestellten Darstellungsmittel befinden sich derzeit bereits im fortgeschrittenen Stadium. Einige Ansätze konnten in dieser Arbeit aufgezeigt werden. Hinsichtlich der Anwendung semantischer Analyseverfahren besteht allerdings derzeit noch reichlich Forschungsbedarf. Nach Ansicht des Autors beschränkt sich der Großteil der Forschung auf diesem Gebiet auf den englischsprachigen Raum.[312]
Neben den bereits in der Arbeit zitierten Instituten stellen das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI GmbH)[313] mit seinen Forschungsbereichen Wissensmanagement und Sprachtechnologie sowie das Institut für Autonome intelligente Systeme (AIS) der Fraunhofer Gesellschaft[314] zwei der wenigen deutschen Konzentrationspunkte dieser Forschungen dar.

312 Eine umfangreiche Liste an Projekten und Publikationen ist z.B. auf http://www.aktors.org zu finden.
313 http://www.dfki.de
314 http://www.ais.fraunhofer.de/de/gf/InfMin.html

Die strukturierte Abbildung von textbasiertem Wissen, d.h. die Überführung von textuellem Wissen in Maschinensprache stößt derzeit schnell an seine Grenzen, wenn man die verschiedenen Definitionen und Anwendungsbereiche (Kontexte) von Wissen berücksichtigen möchte. Die vorgestellte enge Definition von Wissen erlaubt zwar eine Abbildung einfachen Wissens z.B. in Ontologien. Ihr Anwendungsbereich beschränkt sich derzeit jedoch auf die Darstellung von Metawissen, d.h. sie werden nicht zur Wissenskonversion eingesetzt. Mit zunehmender Komplexität und Größe der Ontologien könnte sich dies in Zukunft jedoch ändern. An der Entwicklung von domänenspezifischen sowie abstrakten Ontologien wird derzeit gearbeitet.

Für die Nutzung von Wissen müssen neben einem adäquaten Metainformationssystem auch Schnittstellen definiert werden, über die neue Informationen bzw. neues Wissen in das System fließen kann.[315] Über diese Schnittstellen wäre dann eine sog. *Peer2Peer Content Syndication* realisierbar, bei dem die Beteiligten automatisch von dem "Wissenspool" profitieren könnten.[316] Standards für solche Schnittstellen wurden in Kapitel 6 und 7 vorgestellt.

Damit *Agentensysteme* nützliches Wissen zutage fördern können, sind sie auf die Kommunikation mit anderen Agenten sowie auf aktuelle und zuverlässige Quellen angewiesen. Sie müssen frühzeitig neue, redundanzfreie Informationen liefern und diese nach Relevanz, Auffälligkeit und Signifikanz einordnen bzw. sortieren können. Einige der bereits bestehenden Retrievalsysteme leisten bereits diese Mehrwertdienste. Der sog. *Knowledge Engineer* bzw. Benutzer kann das Wissen nur sinnvoll anwenden, wenn er es in sein bestehendes Anwendungsumfeld bzw. seine Prozesse integrieren kann. Hierfür müssen die Interessen der Benutzer und des Unternehmens ermittelt werden[317] und entsprechende *Benutzerprofile* erstellt werden. Auf Basis dieser Profile können die Agentendienste automatisiert Informationen liefern und beispielsweise bei Eintreffen eines neuen relevanten Dokumentes eine Email versenden. Diese Dienste werden auch *Selective Dissemination of Information* (SDI)- oder *Pushdienste* genannt.[318] Damit die Arbeit der Agentensysteme transparent bleibt bzw. wird, müssen die im Hintergrund ablaufenden Kontext-, Ähnlichkeits- und Relevanzanalysen dem Benutzer visuell präsentiert werden.

315 Hiermit sind sowohl Benutzerschnittstellen als auch die Schnittstellen von Agentensoftware gemeint.
316 [Nejdl2003], [Broekstra2003]
317 [Nakhaeizadeh1998], S.44
318 [Stock2000a], S.117,S.53

Der letzte notwendige Schritt bei der Verwendung des im Semantic Web verfügbaren Wissens besteht in der Schaffung von Vertrauen in die verwendeten Quellen und Dienste. Hierfür ist sicherlich notwendig, dass verschiedene Quellen miteinander abgeglichen, d.h. analysiert und von Instanzen bewertet werden, die bereits Vertrauen bzw. Autorität besitzen (sog. *Trust Center*). Unsicheres Wissen muss als solches erkennbar bzw. entsprechend qualifiziert werden.[319] Nach T. Hoeren sei die Übermittlung von Vertrauen auf elektronischem Weg (fast) nicht möglich:

> "Technik kann niemals Technik legitimieren. Deshalb erweist sich die Frage nach *Trust*, dem Vertrauen in die Integrität und Authentizität elektronischer Texte, als fast unlösbar."[320]

Dass jedoch auch Experten ohne Kenntnisse über die Funktionsweise der Maschinen bereit sind, diesen bereitwillig zu vertrauen, zeigen R. Kuhlens Ausführungen zu J. Weizenbaums Experimente aus dem Jahre 1966:

> "Nie hatte er [J.Weizenbaum] auch nur in Erwägung gezogen, dass nach Berichten über *Eliza* selbst einige Psychiater 1966 die Übertragung des teilnehmenden Verstehens auf Maschinen für möglich und für bald wahrscheinlich hielten. Noch weniger hatte er vermutet – was aber faktisch geschah -, dass Menschen zu diesem Computer eine emotionale, vertrauensvolle Beziehung aufbauen würden."[321]

Da die Verantwortung allerdings letztlich beim Menschen bleibt und nicht auf die Maschinen übertragen werden kann, bliebe nach R. Kuhlen die Bewertung des Wissens letztlich Aufgabe des kompetenten Anwenders:

> "Ein Mensch in der Informationsgesellschaft hat Chancen, ein autonomes, d.h. selbstbestimmtes Individuum zu werden, wenn er informationskompetent ist. Dieser braucht vor den Konsequenzen der technischen Informationsassistenten nicht bange zu sein."[322]

Von dieser Aufgabe entbindet ihn auch nicht das Semantic Web:

> "We are forming cells within a global brain and we are excited that we might start to think collectively. What becomes of us still hangs crucially on how we think individually."[323]

---

319 Unsicheres Wissen (bei dem eindeutige Wahrheitswerte nicht verwendet werden können) kann z.B. durch Häufigkeiten, Wahrscheinlichkeiten, Belieffunktionen, Fuzzy Logic oder temporalen Logiken ("oft, manchmal, immer, nie") beschrieben werden. vgl. [Lusti1990]

320 [Hoeren1998]

321 [Kuhlen1999], S.206

322 Ebd. S.382

323 [Berners-Lee1997]

10 Anhang

10.1 Text Mining Tools

Diese Liste ist eine Zusammenstellung der in der Arbeit erwähnten Anwendungen und Hilfsmittel. Sie sind entsprechend ihres Auftretens in der Arbeit angeordnet.

Content Management Systeme
- Wiki: http://wiki.org
- Twiki http://twiki.org

Knowledgebases
- Cyc Inc., OpenCyc: http://www.opencyc.org, http://www.cyc.com
- MIT, OpenMind,: http://commonsense.media.mit.edu
- Stanford, TAP: http://tap.semanticweb.org, http://tap.stanford.edu/tapkb/

Information Extraction
- ANNIE: http://gate.ac.uk/annie/index.jsp

Knowledge Extraction
- Natural Language Toolkit: http://nltk.sourceforge.net
- Megaputer, TextAnalyst, PolyAnalyst: http://www.megaputer.com/products/

Klassifikationen
- NACE: http://www.fifoost.org/database/nace/
- eCl@ss: http://www.eclass.de
- DDC: http://www.oclc.org/dewey/, http://www.ddc-deutsch.de
- UDC: http://www.udcc.org

Thesauri
- Openthesaurus: http://www.openthesaurus.de, http://openthesaurus.sf.net
- Standard Thesaurus Wirtschaft: http://www.gbi.de/thesaurus/

Topic Maps
- Mindjet, MindManager: http://www.mindjet.com

Ontologien
- SUMO: http://ontology.teknowledge.com
- Dun & Bradstreet: http://www.ksl.stanford.edu/projects/DAML/UNSPSC.daml
- DAML Verzeichnis: http://www.daml.org/ontologies/
- WordNet als OWL: http://taurus.unine.ch/GroupHome/knowler/wordnet.html

Ontologietools
- KAON: http://sourceforge.net/projects/kaon/
- TextToOnto: http://sourceforge.net/projects/texttoonto/
- Protégé, http://protege.stanford.edu
- Chimaera: http://www.ksl.stanford.edu/software/chimaera/
- OntoSQL: http://www.aifb.uni-karlsruhe.de/WBS/aeb/ontosql/
- OntoAgent: http://www.i-u.de/schools/eberhart/ontoagent/
- OntoView http://ontoview.org
- Ontoprise (diverse): http://www.ontoprise.de

Transfer
- SCORM Conformance Test Suite:
  http://www.adlnet.org/index.cfm?fuseaction=SCORDown&listing=Software
- Mindswap, OntoLink:
  http://www.mindswap.org/2004/OntoLink/

Integration
- Teknowledge: SemanticWord, Briefing Associate
  http://mr.teknowledge.com/DAML
- WEKA:
  http://sourceforge.net/projects/weka/, http://www.cs.waikato.ac.nz/~ml/
- GATE:
  http://gate.ac.uk
- Intelligent Miner for Text, IBM:
  http://www.ibm.com/software/data/iminer/fortext/
- SAS®, Text Miner:
  http://www.sas.com/technologies/analytics/datamining/textminer/
- SPSS, Text Mining for Clementine®:
  http://www.spss.com/lexiquest/text_mining_for_clementine.htm
- g.a.d.t GmbH:
  http://gadt.de
- moresophy GmbH, L4:
  http://www.moresophy.de

10.2 DAML-OWL-Beispiel

Um ein Beispiel für in Ontologien abgebildetes Wissen zu geben, werden im Folgenden zwei mögliche Schreibweisen desselben Zusammenhangs gegeben. Beide drücken aus, dass jedes Betriebssystem Software darstellt und das Betriebssystem Linux kostenlos ist.

DAML

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
   xmlns:ns0="http://stefkoch.de/diplom/linux.daml"
   xmlns:xsd="http://www.w3.org/2000/10/XMLSchema#"
   xml:base="http://stefkoch.de/diplom/">

   <daml:Class rdf:about="linux.daml#Software">
      <rdfs:label>Software</rdfs:label>
   </daml:Class>
   <daml:Class rdf:about="linux.daml#Betriebssystem">
      <rdfs:label>Betriebssystem</rdfs:label>
      <rdfs:subClassOf>
         <daml:Class rdf:about="linux.daml#Software"/>
      </rdfs:subClassOf>
   </daml:Class>

   <daml:ObjectProperty rdf:about="linux.daml#kostenlos">
      <rdfs:label>kostenlos</rdfs:label>
   </daml:ObjectProperty>

   <rdf:Description rdf:about="linux.daml#Linux">
      <rdf:type>
         <daml:Class rdf:about="linux.daml#Betriebssystem"/>
      </rdf:type>
      <ns0:kostenlos rdf:resource="linux.daml#Linux"/>
   </rdf:Description>

</rdf:RDF>
```

OWL

```
<?xml version="1.0" encoding="UTF-8"?>
<owls:Ontology xmlns:owls="http://www.w3.org/2003/OWL-XMLSchema"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xml:base="http://stefkoch.de/diplom/">

   <owls:Class owls:name="linux.owl#Software"/>
   <owls:Class owls:name="linux.owl#Betriebssystem"/>

   <owls:SubClassOf>
      <owls:super>
         <owls:Class owls:name="linux.owl#Software"/>
      </owls:super>
      <owls:sub>
         <owls:Class owls:name="linux.owl#Betriebssystem"/>
      </owls:sub>
   </owls:SubClassOf>

   <owls:ObjectProperty owls:name="linux.owl#kostenlos"/>

   <owls:Individual owls:name="linux.owl#Linux">
      <owls:type owls:name="linux.owl#Betriebssystem"/>
      <owls:ObjectPropertyValue owls:property="linux.owl#kostenlos"/>
   </owls:Individual>

</owls:Ontology>
```

Der wesentliche Unterschied zwischen beiden Schreibweisen ist, dass in OWL die Hierarchiebeziehungen von Klassen separat in einem "owls:SubClassOf"-Tag und Instanzen in einem "owls:Individual"-Tag definiert werden können, während DAML hierfür auf die RDF-Definitionen zurückgreift. Die Zuweisung von Properties muss damit bei DAML über den XML-Namespace erfolgen, während bei OWL "owls:ObjectPropertyValue" verwendet werden kann.

Abkürzungen

ACM – Association for Computing Machinery
ADL – Advanced Distributed Learning Initiative
AI – Artificial Intelligence, auch KI
AICC – Aviation Industry CBT Committee
AIEE – American Institute of Electrical Engineers
ARPA – Advanced Research Projects Agency
ASCII – American Standard Code for Information Interchange
CBT – Computer Based Training
CMS – Content Management System
DAML – Darpa Agent Markup Language
DARPA – Defense Advanced Research Projects Agency
DBMS – Datenbank Management System
DDC – Dewey Decimal Classification
DTD – Document Type Definition
EML – Educational Modeling Language
ERP – Enterprise Resource Planning
FAQ – Frequently Asked Questions
HTML – HyperText Markup Language
ISIC – International Standard Industrial Classification of all Economic Activities
KD – Knowledge Discovery
KDD – Knowledge Discovery in Databases
KI – Künstliche Intelligenz, auch AI
KIF – Knowledge Interchange Format
KML – Knowledge Modelling Language
KQML – Knowledge Query and Manipulation Language
KRL – Knowledge Representation Language
KRRS – Knowledge Representation and Reasoning System
KRS – Knowledge Representation System
MIS – Management Information System
MIT – Massachusetts Institute of Technology
NACE – Wirtschaftszweigklassifikation der Europäischen Union
OCR – Optical Character Recognition
OIL – Ontology Inference Language
OKBC – Open Knowledge Base Connectivity
OLAP – OnLine Analytical Processing
OML – Ontology Modelling Language
OSD – Office of the Secretary of Defense
RDF – Resource Description Framework
SCORM – Sharable Content Object Reference Model
SHOE – Simple HTML Ontology Extensions
SQL – Structured Query Language
SWRL – Semantic Web Rule Language
TM – Text Mining
TREC – Text Retrieval Conferences
UDC – Universal Decimal Classification
W3C – World Wide Web Consortium
WBT – Web Based Training
WWW – World Wide Web
OWL – Web Ontology Language
OML – Ontology Modelling Language
XML – eXtensible Markup Language
XOL – XML Ontology Exchange Language
XQL – XML Query Language
XIRQL – XML Information Retrieval Language
XML – eXtensible Markup Language
XSD – XML Schema Definition
XXL – FleXible XML Search Language

Abbildungen

Literatur

- [Aamodt1995]: Agnar Aamodt, Mads Nygård, "Different roles and mutual dependencies of data, information, and knowledge - an AI perspective on their integration", Data Knowledge Engineering, North-Holland Elsevier, Vol.16, 1995, S.191-222, http://citeseer.ist.psu.edu/aamodt95different.html.
- [Abiteboul1997]: Serge Abiteboul, Dallan Quass, Jason McHugh, Jennifer Widom, Janet L. Wiener, "The Lorel query language for semistructured data", International Journal on Digital Libraries, Vol. 1, 1997, http://citeseer.ist.psu.edu/abiteboul97lorel.html.
- [Agrawal1995]: Rakesh Agrawal, Giuseppe Psaila, "Active Data Mining", 1st International Conference on Knowledge Discovery and Data Mining (KDD-95), http://citeseer.ist.psu.edu/agrawal95active.html, S.1.
- [Ahonen-Myka2002]: Helena Ahonen-Myka, "Discovery of Frequent Word Sequences in Text", The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining, Imperial College, London, 2002, http://citeseer.ist.psu.edu/534626.html.
- [Ahonen1997]: Helena Ahonen, Oskari Heinonen, Mika Klemettinen, A. Inkeri Verkamo, "Applying Data Mining Techniques in Text Analysis", 1997, http://citeseer.ist.psu.edu/ahonen97applying.html, S. 9ff..
- [Ahonen1998]: Helena Ahonen, Oskari Heinonen, Mika Klemettinen, A. Inkeri Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections", Advances in Digital Libraries, 1998, http://citeseer.ist.psu.edu/ahonen98applying.html, S.2-11.
- [Aitken2000]: Stuart Aitken, Sandy Reid, "Evaluation of an Ontology-Based Information Retrieval Tool", Workshop on the Applications of Ontologies and Problem-Solving Methods, European Conference on Artificial Intelligence, Berlin, 2000, http://www.aiai.ed.ac.uk/~stuart/Papers/ontologyeval.pdf.
- [Al-Khalifa2003]: Shurug Al-Khalifa, Cong Yu, H. V. Jagadish, "Querying Structured Text in an XML Database", SIGMOD, 2003, http://citeseer.ist.psu.edu/alkhalifa03querying.html.
- [Ankerst2000]: Mihael Ankerst, Martin Ester, Hans-Peter Kriegel, "Towards an Effective Cooperation of the Computer and the User for Classification", Proceedings of the 6th Int. Conference on Knowledge Discovery and Data Mining (KDD) 2000, Boston, MA, http://citeseer.ist.psu.edu/ankerst00towards.html.
- [Baeza-Yates1999]: Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "Modern Information Retrieval", ACM Press, Addison Wesley, New York, 1999, S.124ff..
- [Beier2003]: H. Beier, "Intelligente Informationsstrukturierung und TextMining mit Semantischen Netzen. Intelligent information structuring and text mining with semantic networks", Proceedings of Competence in Content: 25. Online-Tagung der DGI, Hrsg.: R.

Schmidt, Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis (DGI), Frankfurt am Main, DE, 2003, S.78-87, Quelle: INFODATA, FIZ Technik.

- [Berners-Lee1997]: Tim Berners-Lee, "Realising the Full Potential of the Web", Based on a talk presented at the W3C meeting, London, http://www.w3.org/1998/02/Potential.html
- [Bernatzki1996] A. Bernatzki, W. Eppler, H. Gemmeke, "Interpretation of Neural Networks for Classification Tasks", Proceedings of EUFIT 1996, Aachen, Germany, http://citeseer.ist.psu.edu/903.html.
- [Besançon1998]: Martin Rajman, Romaric Besançon, "Text Mining - Knowledge extraction from unstructured textual data", 6th Conference of International Federation of Classification Societies (IFCS-98), Rome, 1998, http://citeseer.ist.psu.edu/besanon98text.html.
- [Bobrov1977]: Daniel G. Bobrov, Terry Winograd, "An Overview of KRL, a knowledge Representation Language.", Cognitive Science, 1977, S. 46ff. ftp://reports.stanford.edu/pub/cstr/reports/cs/tr/76/581/CS-TR-76-581.pdf.
- [Bollacker1998]: Kurt D. Bollacker, Steve Lawrence, C. Lee Giles, "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications", Proceedings of the Second International Conference on Autonomous Agents, 1998, http://citeseer.ist.psu.edu/bollacker98citeseer.html.
- [Borkar2001]: Vinayak Borkar, Kaustubh Deshmukh, Sunita Sarawagi, "Automatic segmentation of text into structured records", Indian Institute of Technology -(DATAMOLD-System), Bombay, ACM SIGMOD 2001, http://citeseer.ist.psu.edu/borkar01automatic.html, S.175.
- [Broekstra2003]: Jeen Broekstra, Marc Ehrig, Peter Haase, Frank van Harmelen, Arjohn Kampman, Marta Sabou, Ronny Siebes, Steen Staab, Heiner Stuckenschmidt, Christoph Tempich, "A Metadata Model for Semantics-Based Peer-To-Peer Systems ", 2003 http://citeseer.ist.psu.edu/584933.html.
- [Callan2003]: Robert Callan, "Neuronale Netze im Klartext", München, Pearson, 2003, S.27.
- [Cerf1969]: Vint Cerf, "Requiem for the Arpanet", http://www.etext.org/Politics/Essays/arpanet.
- [Chamberlin2000]: Don Chamberlin, Jonathan Robie, Daniela Florescu, "Quilt: An XML Query Language for Heterogeneous Data Sources", Lecture Notes in Computer Science, IBM Almaden Research Center, 2000, http://citeseer.ist.psu.edu/chamberlin00quilt.html.
- [Chaudhri1998a]: Vinay K. Chaudhri, Adam Farquhar, Richard Fikes, Peter D. Karp, James P. Rice, "Open Knowledge Base Connectivity 2.0.3 Proposed ", Artificial Intelligence Center SRI International, Knowledge Systems Laboratory Stanford University, 1998, http://www-ksl-svc.stanford.edu:5915/doc/release/okbc/okbc-spec/okbc-2-0-3.pdf, S.1.
- [Chaudhri1998b]: Vinay K. Chaudhri, Adam Farquhar, Richard Fikes, Peter D. Karp, James P. Rice, "OKBC: A programmatic foundation for knowledge base interoperability", Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98), S.600-607, 1998, http://citeseer.ist.psu.edu/chaudhri98okbc.html.
- [Codd1993]: E.F. Codd, S.B. Codd, C.T. Sally, "Providing OLAP (On-Line Analytical Processing) to User-Analysts - an IT mandat.", White paper E.F.,Codd & Associates, 1993.
- [Compton1996]: P. Compton, P. Preston, G. Edwards, B. Kang, "Knowledge Based Systems That Have Some Idea of Their Limits", 1996,

http://citeseer.ist.psu.edu/compton96knowledge.html.

- [Corcho2000a]: Oscar Corcho, Asunción Gómez-Pérez, "A Roadmap to Ontology Specification Languages", Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'00), Juan-les-Pins France, October 2000, http://delicias.dia.fi.upm.es/articulos/ocorcho/ekaw2000-corcho.pdf.
- [Corcho2000b]: Oscar Corcho, Asunción Gómez-Pérez, "Evaluating Knowledge Representation and Reasoning Capabilities of Ontology Specification Languages", Proceedings of the ECAI'00 Workshop on Applications of Ontologies and Problem Solving Methods, Berlin Germany, 2000, http://citeseer.ist.psu.edu/corcho00evaluating.html, S.5.
- [Cui2003]: Z. Cui, J. W. Shepherdson, Y. Li, "An ontology-based approach to eCatalogue management", BT Technology Journal, Vol 21 No 4, October 2003, http://www.kluweronline.com/article.asp?PIPS=5254790.
- [Dasigi1996]: Venu Dasigi, Reinhold Mann, "Neural Net Learning Issues in Classification of Free Text Documents ", AAAI Spring Symposium on Machine Learning in Information Access Technical Papers, 1996, http://citeseer.ist.psu.edu/dasigi96neural.html.
- [Davenport1998]: Thomas Davenport, Laurenc Prusak, "Wenn ihr Unternehmen wüßte, was es alles weiß. Das Praxisbuch zum Wissensmanagement", Landsberg/Lech., 1998. S.186.
- [Davis1993]: Randall Davis, Howard Shrobe, Peter Szolovits, "What is a Knowledge Representation?", AI Magazine, 14, S.17-33, MIT AI Lab, 1993, http://medg.lcs.mit.edu/ftp/psz/aimag-final.ps.
- [DeBra1994]: Paul De Bra, Geert-Jan Houben, Joep De Vocht, Yoram Kornatzky, "Retrieval of Hypertext Structures", Proceedings of Stinfon-94 Conference, Tilburg, 1994, http://citeseer.ist.psu.edu/108018.html.
- [Deutsch1998]: Alin Deutsch, Mary Fernandez, Daniela Florescu, Alon Levy, Dan Suciu, "XML-QL: A Query Language for XML", Proceedings of WWW The Query Language Workshop QL, Cambridge, MA, 1998, http://citeseer.ist.psu.edu/390950.html.
- [Desmontils2001]: Emmanuel Desmontils, Christine Jacquin, "Indexing a Web Site with a Terminology Oriented Ontology", Proceedings of SWWS'01, The first Semantic Web Working Symposium, Stanford University, 2001, http://citeseer.ist.psu.edu/et02indexing.html, S.7.
- [Ding2002]: Anne Denton, Qiang Ding, Qin Ding, William Perrizo, "Efficient Hierarchical Clustering of Large Data Sets Using P-trees", Proceedings of 15th International Conference on Computer Applications in Industry and Engineering (CAINE'02), San Diego, CA, Nov. 2002, S. 138-141, http://cs.hbg.psu.edu/~ding/publications/CAINE_109.pdf.
- [Dixon1997]: Mark Dixon, "An Overview of Document Mining Technology", 1997, http://citeseer.ist.psu.edu/dixon97overview.html, S.1.
- [Doan2002]: An Hai Doan, Jayant Madhavan, Pedro Domingos, Alon Halevy, "Learning to Map between Ontologies on the Semantic Web." , Proceedings of the 11th International World Wide Web Conference (S.662-673), 2002. Honolulu, ACM Press, http://www.cs.washington.edu/homes/pedrod/papers/www02.pdf, S.4.
- [Eberhart2002]: Andreas Eberhart, "OntoAgent: A Platform for the Declarative Specification of Agents", Proceedings of the ISWC 2002 Rule Markup Languages for Business Rules on the Semantic Web, http://citeseer.ist.psu.edu/eberhart02ontoagent.html.

- [Ester1998]: Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, "Clustering for Mining in Large Spatial Databases", Special Issue on Data Mining, KI-Journal, ScienTec Publishing, Vol. 1, 1998, http://www.cs.helsinki.fi/u/gionis/seminar_papers/ester98clustering.pdf., S.2
- [Ester2000]: Mihael Ankerst, Martin Ester, Hans-Peter Kriegel, "Cooperative Classification: A Visualization-Based Approach of Combining the Strengths of the User and the Computer", Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers. http://citeseer.ist.psu.edu/455891.html.
- [Fayyad1996a]: Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communication of the ACM, Vol.29, 1996, http://citeseer.ist.psu.edu/fayyad96kdd.html.
- [Fayyad1996b]: Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, American Association for Artificial Intelligence, 0738-4602-1996, 1996, S.39, http://citeseer.ist.psu.edu/fayyad96from.html.
- [Feldman1998]: Ronen Feldman, Moshe Fresko, Yakkov Kinar, Yehuda Lindell, Orly Liphstat, Martin Rajman , Yonatan Schler, Oren Zamir, "Text Mining at the Term Level", Principles of Data Mining and Knowledge Discovery, 1998, http://citeseer.ist.psu.edu/feldman98text.html.
- [Fensel2000]: Dieter Fensel, Ian Horrocks, Frank van Harmelen, Stefan Decker, Michael Erdmann, and Michel C. A. Klein, "OIL in a Nutshell", Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling, and Management (EKAW'00)", Springer, 2000, http://www.cs.vu.nl/~ontoknow/oil/downl/oilnutshell.pdf.
- [Fensel2001]: Dieter Fensel, "Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce.", Springer-Verlag, 2001, http://citeseer.ist.psu.edu/413498.html.
- [Franke2003]: Ingrid Renz, Jürgen Franke, "Text Mining", Text Mining - Theoretical Aspects and Applications, Physica, 2003, S.1f.
- [Freitag1998]: Dayne Freitag, "Information Extraction from HTML: Application of a General Machine Learning Approach", AAAI, 1998, http://citeseer.ist.psu.edu/freitag98information.html.
- [Fuhr2001]: Norbert Fuhr, Kai Großjohann, "XIRQL: A Query Language for Information Retrieval in XML Documents", 2001 http://citeseer.ist.psu.edu/fuhr01xirql.html.
- [Fuhr2002]: Mohammad Abolhassani, Norbert Fuhr, Norbert Gövert, Kai Großjohann, "HyREX: Hypermedia Retrieval Engine for XML", Research Report an der University of Dortmund, Department of Computer Science, 2002, http://www.is.informatik.uni-duisburg.de/bib/xml/Fuhr_etal_02b.html.
- [Fuhr2003]: Norbert Fuhr, Kai Großjohann, S. Kriewel, "A Query Language and User Interface for XML Information Retrieval", Intelligent XML Retrieval Vol. 2818, Springer, 2003, http://www.is.informatik.uni-duisburg.de/bib/xml/Fuhr_etal_03a.html.
- [Fuhr2004]: Norbert Fuhr, Kai Großjohann, "XIRQL: An XML Query Language Based on Information Retrieval Concepts", ACM Transactions on Information Systems, Volume 22, 2004, S.313—356, http://www.is.informatik.uni-duisburg.de/bib/xml/Fuhr_Grossjohann_04.html, S.36.
- [Gemert2000]: Jan van Gemert, "Text Mining Tools on the Internet - An overview", Intelligent Sensory Information Systems (ISIS), University of Amsterdam, 2000,

http://carol.science.uva.nl/~jvgemert/mia_page/textminingtools.pdf

- [Gil2000]: Yolanda Gil, Varun Ratnakar, "A Comparison of (Semantic) Markup Languages", American Association for Artificial Intelligence, 2000, http://www.isi.edu/expect/web/semanticweb/paper.pdf.
- [Giles1998]: C. Lee Giles, Kurt D. Bollacker, Steve Lawrence, "CiteSeer: An Automatic Citation Indexing System", Digital Libraries 98 - The Third ACM Conference on Digital Libraries, 1998, http://citeseer.ist.psu.edu/108208.html.
- [Gödert1998]: Elisabeth Sachse, Martina Liebig, Winfried Gödert, "Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt.", Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft, Band 14, 1998, http://www.fbi.fh-koeln.de/institut/papers/kabi/volltexte/band014.pdf, S.37.
- [Gómez-Pérez2004]: Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho, "Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web ", Springer Informatik, 2004.
- [Gonzalo1998]: Julio Gonzalo, Felisa Verdejo, Irina Chugur, Juan Cigarran, "Indexing with WordNet synsets can improve Text Retrieval", Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, http://citeseer.ist.psu.edu/gonzalo98indexing.html.
- [Guha2003]: R. Guha, Rob McCool, "TAP: A Semantic Web Platform", Computer Networks: The International Journal of Computer and Telecommunications Networking , Volume 42 , Special issue: The Semantic Web: an evolution for a revolution, 2003, http://tap.stanford.edu/tap.pdf.
- [Hahn1997]: Udo Hahn, Klemens Schnattinger, "A Qualitative Growth Model for Real-World Text Knowledge Bases", RIAO`97 – Proceedings of the 5th Conference on Computer-Assisted Information Searching on the Internet, Montreal, Quebec, Canada, 1997, S.578-597 http://citeseer.ist.psu.edu/hahn97qualitative.html, S.3.
- [Heikkinen2000]: B. Heikkinen, "Generalization of Document Structures and Document Assembly", 2000, http://citeseer.ist.psu.edu/heikkinen00generalization.html, S.1.
- [Hoeren1998]: T. Hoeren, "Internet und Recht – Neue Paradigmen des Informationsrechts.", Neue Juristische Wochenschrift 51, 1998, S.2854, aus: [Kuhlen1999], S.367.
- [Hönig1998]: Thomas Hönig, "Data Warehousing, Data Mining OLAP", Hrsg.: Wolfgang Martin, International Thomson Publishing, Bonn, 1998.
- [Honkela1997]: Timo Honkela "WEBSOM Self-Organizing Maps of Document Collections" , Proceedings of Workshop on Self-Organizing Maps WSOM'97, Espoo, Finland, http://citeseer.ist.psu.edu/honkela97websom.html.
- [Horrocks2000]: Ian Horrocks, "A Denotational Semantics for OIL-Lite and Standard Oil", Department of Computer Science University of Manchester, UK, 2000, http://citeseer.ist.psu.edu/337591.html.
- [Hotho2002]: Andreas Hotho, Alexander Maedche, Steffen Staab, "Ontology-based Text Document Clustering", http://www.aifb.uni-karlsruhe.de/WBS/aho/pub/Ontology_based_Text_Document_Clustering_2002.pdf.
- [Hotho2003a]: Andreas Hotho, Steffen Staab, Gerd Stumme, "WordNet improves text document clustering", Proceedings of the SIGIR 2003 Semantic Web Workshop,

http://www.aifb.uni-karlsruhe.de/WBS/aho/pub/hothoetal_sigir_ws_sem_web.pdf

- [Hotho2003b]: Andreas Hotho, Steffen Staab, Gerd Stumme, "Ontologies Improve Text Document Clustering", Proceedings of the ICDM 03, The 2003 IEEE International Conference on Data Mining, http://www.aifb.uni-karlsruhe.de/WBS/aho/pub/hothoa_icdm_poster03.pdf.
- [Kalfoglou2002]: Yannis Kalfoglou, Harith Alani, Kieron O'Hara, Nigel Shadbolt, "Initiating Organizational Memories using Ontology Network Analysis.", Knowledge Management and Organizational Memories workshop, 15th European Conference on Artificial Intelligence, Lyon, France, 2002, http://citeseer.ist.psu.edu/kalfoglou02initiating.html.
- [Karvounarakis2002]: Greg Karvounarakis, Sofia Alexaki, Vassilis Christophides, Dimitris Plexousakis, Michel Scholl, "RQL: A Declarative Query Language for RDF", The 11th Intl. World Wide Web Conference (WWW2002), http://www.ai.mit.edu/people/jimmylin/papers/Karvounarakis02.pdf
- [Kifer1990]: Michael Kifer, Georg Lausen, "F-Logic: A Higher-Order Language for Reasoning about Objects, Inheritance, and Scheme", 1990, http://citeseer.ist.psu.edu/kifer90flogic.html.
- [Kirchner1998]: Joachim Kirchner, "Data Warehousing, Data Mining- OLAP", Hrsg.: Wolfgang Martin, International Thomson Publishing, Bonn, 1998, S. 151.
- [Kirschner2003]: Paul A. Kirschner, 2003 , "Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making.", http://www.visualizingargumentation.info
- [Klein2002]: Michel Klein, Dieter Fensel, Atanas Kiryakov, Damyan Ognyanov, "Ontology versioning and change detection on the Web", 2002, http://citeseer.ist.psu.edu/klein02ontology.html.
- [Kohonen2001]: T. Kohonen, "Self-Organizing Maps", Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 2001.
- [König1998]: Andreas König, "A Survey of Methods for Multivariate Data Projection, Visualisation and Interactive Analysis", Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems IIZUKA'98, S. 55-59, Iizuka, Fukuoka, Japan, October 1998. http://citeseer.ist.psu.edu/87942.html.
- [Kosala2002]: Raymond Kosala, Jan Van den Bussche, Maurice Bruynooghe, Hendrik Blockeel, "Information extraction in structured documents using tree automata induction.", Proceedings of the the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), 2002, http://citeseer.ist.psu.edu/article/kosala02information.html.
- [Kuhlen1999]: Rainer Kuhlen, "Die Konsequenzen von Informationsassistenzen – Was bedeutet informationelle Autonomie oder wie kann Vertrauen in elektronische Dienste in offenen Informationsmärkten gesichert werden?", Suhrkamp, Frankfurt am Main, 1999
- [Kurz1998]: Andreas Kurz, "Data Warehousing, Data Mining OLAP", Hrsg.: Wolfgang Martin, International Thomson Publishing, Bonn, 1998, S. 252.
- [Kushmerick2001]: Nicholas Kushmerick, Edward Johnston, Stephen McGuinness, "Information Extraction By Text Classification", 2001, http://citeseer.ist.psu.edu/kushmerick01information.html.
- [Labrou1999]: Yannis Labrou, Tim Finin, "Yahoo! as an Ontology Using Yahoo! Categories to Describe Documents", CIKM, 1999, http://citeseer.ist.psu.edu/labrou99yahoo.html.
- [Langley1992]: Pat Langley, Wayne Iba, Kevin Thompson, "An Analysis of Bayesian

Classifiers", National Conference on Artificial Intelligence, NASA Ames Research Center, 1992, http://citeseer.ist.psu.edu/langley92analysis.html, S. 223-228.

- [Lanquillon2001]: Carsten Lanquillon, "Enhancing Text Classification to Improve Information Filtering", Dissertation an der Uni Magdeburg, 2001, http://citeseer.ist.psu.edu/lanquillon01enhancing.html.
- [Lawrence1999]: Steve Lawrence, C. Lee Giles, Kurt Bollacker, "Digital Libraries and Autonomous Citation Indexing", IEEE Computer, Volume 32, Number 6, 1999, http://citeseer.ist.psu.edu/lawrence99digital.html, S.67-71.
- [Lem1964]: Stanislaw Lem, "Summa technologiae", Suhrkamp, S.140.
- [Lepsky1997]: Klaus Lepsky, "Auf dem Weg zur automatischen Inhaltserschließung? Das DFG-Projekt MILOS und seine Ergebnisse.", Mitteilungen der Gesellschaft für Bibliothekswesen und Dokumentation des Landbaues, Heft 53, 1997, S.46-52.
- [Lepsky1998]: Winfried Gödert, Klaus Lepsky, "Semantische Umfeldsuche im Information Retrieval.", Zeitschrift für Bibliothekswesen und Bibliographie 45, Heft 4, 1998, S.401-423.
- [Lin2001]: Dekang Lin, Patrick Pantel, "DIRT Discovery of Inference Rules from Text", Knowledge Discovery and Data Mining, 2001, http://citeseer.ist.psu.edu/lin01dirt.html.
- [Lin2002]: Dekang Lin, Patrick Pantel, "Concept Discovery from Text", Department of Computing Science University of Alberta, 2002, http://citeseer.ist.psu.edu/lin02concept.html.
- [Loh2003]: Stanley Loh, José Palazzo M. de Oliveira, Mauricio A. Gameiro, "Knowledge Discovery in Texts for Constructing Decision Support Systems", Applied Intelligence 18, S. 357-366, Kluwer Academic Publishers, 2003, http://www.kluweronline.com/article.asp?PIPS=5119087.
- [Lugo2002]: Gustavo A. Giménez-Lugo, Analia Amandi, Jaime Sichman, Daniela Godoy, "Enriching Information Agents' Knowledge by Ontology Comparison: A Case Study", 2002, http://citeseer.ist.psu.edu/572772.html.
- [Lusti1990]: Markus Lusti, "Wissensbasierte Systeme", Hrsg.: Karl Heinz Böhling, Mannheim/Wien/Zürich, BI Wissenschaftsverlag, 1990.
- [Martin1998]: Wolfgang Martin, "Data Warehousing, Data Mining - OLAP", Hrsg.: Wolfgang Martin, International Thomson Publishing, Bonn, 1998, S. 418.
- [May2001]: Wolfgang May, "Integration of XML Data in XPathLog", DIWeb, 2001, http://citeseer.ist.psu.edu/may01integration.html.
- [Maynard2003]: Diana Maynard, "Information Extraction - why Google doesn't even come close ", Natural Language Processing Group, University of Sheffield, UK, BCS meeting, 25 September 2003, http://gate.ac.uk/sale/talks/bcs-03-cheltenham.ppt.
- [McEntire2000]: Robin McEntire, Peter Karp, Neil Abernethy, David Benton, Gregg Helt, Matt DeJongh, Robert Kent, Anthony Kosky, Suzanna Lewis, Dan Hodnett, Eric Neumann, Frank Olken, Dhiraj Pathak, Peter Tarczy-Hornoch, Luca Toldo, Thodoros Topaloglou, "An Evaluation of Ontology Exchange Languages for Bioinformatics", 2000, http://citeseer.ist.psu.edu/mcentire00evaluation.html.
- [Michie1994]: Donald Michie, D. J. Spiegelhalter, C. C. Taylor, "Machine Learning, Neural and Statistical Classification", Ellis Horwood, 1994, http://citeseer.ist.psu.edu/michie94machine.html, S.216.
- [Mladenic1998]: Dunja Mladenic, "Turning Yahoo into an Automatic Web-Page Classifier", 13th European Conference on Artificial Intelligence Young Researcher Paper, John Wiley &

Sons, Ltd., 1998, http://citeseer.ist.psu.edu/mladenic98turning.html.

- [Mooney2002]: U. Y. Nahm, R. J. Mooney, "Text Mining with Information Extraction", AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002, Department of Computer Sciences, University of Texas, http://www.cs.utexas.edu/users/ml/papers/discotex-aaaisymp-02.pdf, S.2.
- [Mooney2003]: R. J. Mooney, "Intelligent Information Retrieval and Web Search", Kursmaterial University of Texas, 2003, http://www.cs.utexas.edu/users/mooney/ir-course/slides/TextCategorization.ppt, S.16.
- [Müller1998]: Müller, Hausdorf, Schneeberge, "Data Mining", Hrsg.: Gholamreza Nakhaeizadeh, Physica-Verlag, 1998.
- [Myka1992]: Andreas Myka, F. Sarre, Ulrich Güntzer, "Rulebased machine learning of hypertext links", 1992, http://citeseer.ist.psu.edu/myka92rulebased.html.
- [Myka1995]: Andreas Myka, Ulrich Güntzer, "Automatic Hypertext Conversion of Paper Document Collections", 1995, http://citeseer.ist.psu.edu/myka95automatic.html.
- [Myka1996a]: Andreas Myka, Ulrich Güntzer, H. Argenton, "Towards Automatic Hypertextual Representation of Linear Texts", PODP, 1996, http://citeseer.ist.psu.edu/43202.html.
- [Myka1996b]: Andreas Myka, Ulrich Güntzer, "Fuzzy Full-Text Searches in OCR Databases", Advances in Digital Libraries, 1996, http://citeseer.ist.psu.edu/myka96fuzzy.html.
- [Myka1996c]: Andreas Myka, Ulrich Güntzer, "Processing Hypertext Link Descriptions", 1996, http://citeseer.ist.psu.edu/43066.html, S.4.
- [Myka1997]: Andreas Myka, Ulrich Güntzer, "On Automatic Similarity Linking in Digital Libraries", Proceedings of DEXA'97 Workshop", S.278-283, 1997, http://citeseer.ist.psu.edu/myka97automatic.html.
- [Nakhaeizadeh1998]: Gholamreza Nakhaeizadeh, "Data Mining" Hrsg.: Gholamreza Nakhaeizadeh, Physica-Verlag 1998.
- [Nejdl2003]: Wolfgang Nejdl, Martin Wolpers, Wolf Siberski, Christoph Schmitz, Mario Schlosser, Ingo Brunkhorst, Alexander Löser, "Super-Peer-Based Routing and Clustering Strategies for RDF-Based Peer-to-Peer Networks", Proceedings of the 12th International World Wide Web Conference, Budapest, Ungarn, 2003, http://citeseer.ist.psu.edu/nejdl03superpeerbased.html.
- [Nilsson1980]: Nils J. Nilsson, "Principles of Artificial Intelligence", Tioga, Palo Alto, 1980.
- [Noy2000]: Natalya Fridman Noy, Ray W. Fergerson, Mark A. Musen, "The knowledge model of Protégé-2000: combining interoperability and flexibility", http://www-smi.stanford.edu/pubs/SMI_Reports/SMI-2000-0830.pdf.
- [Pantel2002]: Patrick Pantel, Dekang Lin, "Discovering Word Senses from Text", University of Alberta Department of Computing Science Edmonton, Canada, SIGKDD'02, http://citeseer.ist.psu.edu/570332.html.
- [Pinto2004]: So a Pinto, Steffen Staab, York Sure, and Christoph Tempich, "OntoEdit empowering SWAP: A case study in supporting DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies (DILIGENT)", http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/2004_esws_diligent.pdf.
- [Quinlan1986]: J. R. Quinlan, "Induction of Decision Trees. Machine Learning", Machine

Learning, Kluwer Journals, 1986, http://www.kluweronline.com/article.asp?PIPS=422606.

- [Rajman1997]: Martin Rajman, Romaric Besancon, "Text Mining: Natural Language Techniques and Text Mining Applications", Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7), Chapam & Hall IFIP, 1997, S.4, http://citeseer.ist.psu.edu/rajman97text.html.
- [Ray2001]: Erik T. Ray, "Learning XML", 1st Edition, O'Reilly & Associates, 2001.
- [Reategui1997]: Eliseo B. Reategui, John A. Campbell, Beatriz F. Leao, "A Case-Based Model that Integrates Specific and General Knowledge in Reasoning" Applied Intelligence Vol. 7, 1997, Kluwer Academic Publishers, Netherlands, http://www.kluweronline.com/article.asp?PIPS=121979, S.79-90.
- [Rich1983]: Elaine Rich, "Articial Intelligence", McGraw-Hill Book, New York, 1983, S.14.
- [Riloff1994]: Ellen Riloff, Wendy Lehnert, "Information Extraction as a Basis for High-Precision Text Classification", 1994, ACM Transactions on Information Systems, http://citeseer.ist.psu.edu/riloff94information.html, S. 296.
- [Rötzer1999]: Florian Rötzer, "Megamaschine Wissen - Vision: Überleben im Netz", Campus Verlag, Frankfurt, New York, 1999, S.176.
- [Salton1983]: Gerard Salton, Michael J. McGill, "Information Retrieval. Grundlegendes für Informationswissenschaftler. (Originaltitel: 'Introduction to Modern Information Retrieval.')", McGraw-Hill,1983, S.68-69.
- [Savoy1991]: Jacques Savoy, Daniel Desbois, "Information retrieval in hypertext systems: an approach using Bayesian networks", Electronic Publishing/Origination, Dissemination, and Design, 1991, http://citeseer.ist.psu.edu/savoy91information.html.
- [Shannon1980]: Claude F. Shannon, Warren Weaver, "The Mathematical Theory of Communication", University of Illinois Press, U.S., 1980.
- [Sherif2002]: Yacoub Sherif, "Bootstrapping Semantic Web Languages using a UML Meta-Modeling Approach," Information Infrastructure Laboratory HP Laboratories Palo Alto, 2002, http://www.hpl.hp.com/techreports/2002/HPL-2002-200.html.
- [Stenmark2001]: Dick Stenmark, "The Relationship between Information and Knowledge", http://citeseer.ist.psu.edu/stenmark01relationship.html.
- [Stock2000a]: Wolfgang G. Stock, "Informationswirtschaft: Management externen Wissens", Oldenbourg, München/Wien, 2000.
- [Stock2000b]: Wolfgang G. Stock, "Textwortmethode", PASSWORD 07 und 08/2000, http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/1/1078740450password_7.pdf.
- [Su2002]: Xiaomeng Su, Lars Ilebrekke, "A Comparative Study of Ontology Languages and Tools", Hrsg.: A. Banks Pidduck, Springer-Verlag, Berlin/Heidelberg, 2002, http://citeseer.ist.psu.edu/556209.html.
- [Sure1999]: York Sure, Rudi Studer, "On-To-Knowledge Methodology - Employed and Evaluated Version", University of Karlsruhe On-To-Knowledge EU IST-1999-10132 Project Deliverable D16 (WP5), http://citeseer.ist.psu.edu/sure99toknowledge.html, S.48-49.
- [Sure2003]: York Sure, J. Angele, S. Staab, "OntoEdit: Multifaceted inferencing for ontology engineering", Journal on Data Semantics, LNCS 2800, 2003, S.128-152, http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/ontoedit-data-semantics.pdf.

- [Sure2004]: Marc Ehrig, York Sure, "Ontology Mapping - An Integrated Approach", Institute AIFB, University of Karlsruhe, 2004, http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/2004_esws_mapping.pdf.
- [Takeda1995]: Hideaki Takeda, Kenji Iino, Toyoaki Nishida, "Agent organization and communication with multiple ontologies.", International Journal of Cooperative Information Systems, 4, 1995, S.321-337, http://ai-www.aist-nara.ac.jp/papers/takeda/ps/ijicis.ps.gz.
- [Theobald2002a]: Anja Theobald, Gerhard Weikum, "The XXL Search Engine: Ranked Retrieval of XML Data Using Indexes and Ontologies", ACM SIGMOD 2002, http://ranger.uta.edu/~alp/ix/readings/theobaldXXL-sigmod02.pdf.
- [Theobald2002b]: Anja Theobald, Gerhard Weikum, "The index-based XXL search engine for querying XML data with relevance ranking", EDBT, 2002, http://ranger.uta.edu/~alp/ix/readings/theobaldXXL-EDBT02.pdf.
- [Theobald2003]: Anja Theobald, "An Ontology for Domain-oriented Semantic Similarity Search on XML Data", 10th Conference on Database Systems for Business, Technology and Web (BTW), Leipzig, 2003, http://www.mpi-sb.mpg.de/units/ag5/publications/theobald_btw2003.pdf.
- [Ultsch2003]: Alfred Ultsch, "U*-Matrix: a Tool to visualize Clusters in high dimensional Data", 2003, http://www.mathematik.uni-marburg.de/forschung/publikationen/paper_info/bfi36.pdf.
- [Vega1998]: Julio César Arpírez Vega, Asunción Gómez-Pérez, Adolfo Lozano Tello, Helena Sofia Andrade N. P. Pinto, "(ONTO)2Agent: An ontology-based WWW broker to select ontologies", 13th European Conference on Artificial Intelligence ECAI'98, Brighton, England, 1998, http://citeseer.ist.psu.edu/34215.html, S.1, S.7.
- [Vizine-Goetz2001]: Diane Vizine-Goetz, "Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources", Library of Congress, U.S., 2001, http://www.loc.gov/catdir/bibcontrol/vizinegoetz_paper.html.
- [Wedekind1998]: Hartmut Wedekind, "Datenorganisation", 3. Aufl., de Gruyter, 1989.
- [Wilde2001]: Klaus Wilde, "Data Warehouse, OLAP und Data Mining im Marketing", Handbuch Data Mining im Marketing, Wiesbaden: Vieweg, Gabler 2001, S.14
- [Winston1987]: Patrick Henry Winston, "Künstliche Intelligenz", Addison-Wesley, Bonn, 1987, S.21.
- [Wittgenstein1949]: Wittgenstein, Ludwig, "Philosophische Untersuchungen", zitiert aus: Hans Joachim Störig, "Kleine Weltgeschichte der Philosophie", Fischer Taschenbuch Verlag, Frankfurt am Main, 1993, S. 658.
- [Wong1986]: S.K.M. Wong, W. Ziarko, "A machine learning approach to information retrieval" Department of Computer Science, University of Regina, Regina. Saskatchewan. Canada, 1986 ACM Conference on Research and Development in Information Retrieval, http://portal.acm.org/citation.cfm?id=253217.
- [Yang1997]: Yiming Yang, "An Evaluation of Statistical Approaches to Text Categorization", Information Retrieval, Kluwer Academic Publishers, 1997, http://citeseer.ist.psu.edu/yang97evaluation.html.
- [Zavrel2000]: Jakub Zavrel, Peter Berck, Willem Lavrijssen, "Information Extraction by Text Classification: Corpus Mining for Features", LREC2000, http://citeseer.ist.psu.edu/zavrel00information.html.

Eidesstattliche Erklärung

Hiermit versichere ich, die Arbeit selbstständig verfasst
und keine anderen als die angegebenen Quellen
und Hilfsmittel benutzt zu haben.

_____ _____

Ort, Datum Unterschrift