Freie wissenschaftliche Arbeit zur Erlangung
des Grades eines Diplom-Wirtschaftsinformatikers
mit dem Thema:

Evaluierung unterschiedlicher Methoden zur Identifikation von Wissensstrukturen
---
Recherche, Vorstellung, Analyse und Abgrenzung theoretischer Konzepte und Paradigmen

Eingereicht am: 07.03.2002

Eingereicht bei:
Prof. Dr. Ludwig Nastansky
Lehrstuhl für Wirtschaftsinformatik 2
Business Computing 2
Universität Paderborn

Von: Shengxi Long
Aus: VR China
Geb. am 20.08.1968

Inhaltsverzeichnis

1 Einleitung…. 3
2 Wissensstrukturen und Informationsmanagement 5
3 Historische Ansätze 9
3.1 Piaget's Theorie 9
3.2 Churchman's Notation 13
4 Methoden zur Identifikation von Wissensstrukturen 16
4.1 Knowledge Discovery in Database und Data Mining 16
4.1.1 Begriffsdefinitionen 16
4.1.2 Der KDD-Prozeß 18
4.1.3 Data Mining-Methoden 19
4.1.3.1 Aufgaben von Data Mining 19
4.1.3.2 Überblick über die Data Mining-Methoden 21
4.1.4. Data Mining-Techniken 23
4.1.4.1 Entscheidungsbäume 24
4.1.4.2 Neuronale Netze 26
4.1.4.3 Assoziationsregeln 28
4.1.4.4 Clustering 29
4.1.4.5 Statistische Verfahren 30
4.2 OLAP…. 33
4.2.1 Abgrenzung zu OLTP 33
4.2.2 Grundlagen 33
4.3 Information Mapping 37
4.3.1 Einführung 37
4.3.2 Was ist Information Mapping 38
4.3.3 Die Grundelementen der Information-Mapping-Methode 37
4.3.3.1 Informationseinheiten 37
4.3.3.2 Informationstype 38
4.3.3.3 Prinzipien 39
4.4 RDF…. 42
4.4.1 Matadaten 42
4.4.2 RDF-Datenmodell 43
4.4.3 RDF-Syntax 46
4.4.4 RDF-Schema 48
4.4.4.1 Klassen und Eigenschaften 49
4.4.4.2 Beschränkungen 51

4.5 Topic Maps ... 53
4.5.1 Einführung ... 53
4.5.2 Ein Beispiel zur Problemdarstellung ... 53
4.5.3 Grundlegende Begriffserklärung ... 54
4.5.4 ISO-Standard Topic Maps ... 54
4.5.4.1 Topics ... 55
4.5.4.2 Topic Types ... 55
4.5.4.3 Topic Names ... 56
4.5.4.4 Occurrences ... 56
4.5.4.5 Associations ... 58
4.5.4.6 Public Subject Descriptor ... 59
4.5.4.7 Scopes ... 61
4.5.4.8 Facets ... 62
5 Abgrenzung verschiedener Methoden ... 63
5.1 KDD und OLAP ... 63
5.2 Data Mining und OLAP ... 64
5.3 KDD und Data Mining ... 65
5.4 RDF und Topic Maps ... 66
5.5 Information Mapping und Topic Maps ... 66
6 Einsatzmöglichkeiten ... 67
6.1 Data Mining ... 67
6.1.1 Geschäftliche Transaktionen ... 67
6.1.2 E-Commerce ... 68
6.1.3 Webserver und Internetdaten ... 68
6.2 OLAP ... 68
6.3 RDF ... 69
6.4 Topic Maps ... 69
6.4.1 Dokumentenmanagement ... 69
6.4.2 Internet ... 71
7 Zusammenfassung und Ausblick ... 72
8 Literaturverzeichnis ... 76
9. Abkürzungsverzeichnis ... 79
Anhang A ... 80
Anhang B ... 83

1 Einleitung

Die Informationstechnologien haben beigetragen, die Informationen, die zuvor auf Papieren in den Ordnern, Schränken oder Schreibtischen verstreut waren, in elektronische Dokumente zu verwandeln und verteilt zu verwalten. Beispiele für eine solche Speicherung sind vornehmlich in Datenbanken oder elektronischen Ordnern zu finden. Über eine Netzverbindung der einzelnen Arbeitsplatzrechner untereinander und mit Server wurde eine organisationsweite Informationsstruktur geschaffen. Durch die ständigen Entwicklung der Informations- und Kommunikationstechnologie, deren immer stark werdenden Einsätze in betriebliche Prozess wie auch in organisatorische Verwaltung, führt zur immer wachsenden Anzahl an Datenbanken, auf denen die zahlreiche Informationen gespeichert sind. Es bleibt jedoch oft schwierig, die Wissensstrukturen aus den riesigen Datenbeständen zu identifizieren und die gewünschten Informationen wieder zu finden. Ein Grund für das schwere Auffinden der relevanten Informationen ist die oft anzutreffende Unkenntnis in welche Datenbanken nachzusehen ist oder die Nichtwissen der Stelle, an der sich die benötigten Dokumente befinden. So müssen wir viele Zeit damit verbringen, die relevante Informationen zu finden, zu sichten und zu werten. Die US-amerikanische Warenhauskette Wal-Mart, mit mehr als 2000 Geschäften eines der größten Einzelhandelsunternehmen weltweit, speichert mehr als 20 Millionen Transaktionen täglich. Groupware Competence Center(GCC) in der Universität Paderborn hat mehr als 30 Datenbanken eingesetzt, auf denen mehr als 60.000 Dokumenten gespeichert sind, welche die unterschiedliche Kontexte wie Veröffentlichungen, Lehrveranstaltungen und Projekte etc. enthalten. Mit diesen großen Datenmengen ist der Mensch nicht in der Lage, einen Überblick über die Informationenbestände zu behalten. Schätzungsweise sind bis zu 80 Prozent der betrieblichen Informationen in unstrukturierten Dokumenten abgelegt und auch nicht im Wertschöpfungsprozess eingesetzt. Diese riesigen Informationsmenge, deren Analyse und Wissensentdeckung für eine Erzielung von Wettbewerbsvorteilen unumgänglich ist, stellt potentielle Quelle wissensorientierter Informationen dar[vgl. Saü00]. Der Einsatz neuer Informationstechnologien wird immer notwendig. In der Vergangenheit wurden zahlreiche Informationstechnologien entwickelt, die unterschiedliche Mechanismen und Funktionalitäten liefern. wie SQL-Abfrage in Datenbankmanagementsysteme oder Volltextsuchmechanismen in Lotus Notes. Zur Lösung dieser Problemstellung liefern

sie nur wenige funktionelle Unterstützung.

Mit wachsenden organisationalen Wissensbasen auf der einer Seite und dem Mangel an effektiven Mechanismen und Funktionalitäten für die Navigation, die Verknüpfung und das Suchen auf der anderer Seite[Smi01], wird das Bedürfnis nach leistungsfähige Konzepte und erweiterte Mechanismen , die den Anwender bei der Suche nach relevanten Informations- und Wissensobjekt sowie bei der Wissensnutzung und Wissensidentifikation Effektiver unterstützen.

In der vorliegenden Arbeit werden sowohl die historische theoretische Methoden als auch die praxis- und zukunftsorientierte Methoden zur Wissensrepräsentation und Wissensidentifikation vorgestellt, um einen Blick über diese Forschungsgebiet zu gewinnen.

In Kapitel 2 werden die grundlegende Begriffe Wissen und Information erläutert. Anschließend wird ein Überblick über den Aufbau der Wissensstruktur im Zusammenhang mit Informationsmanagement angegeben.

In Kapitel 3 befasst sich mit Piaget's Theorie über die kognitiven Entwicklung des Kinds und Churchman's Notation über die Konstruktion von Informationssystem, die die Wissensentdeckung von Kind wie auch bei Maschinen aufweisen. Bei der Beschreibung dieser Methoden verzichtet ich bewusst auf eine detaillierte Beschreibung der einzelnen Theorien. Die entscheidende Ursache für dieses wird mit den äußerst schwierigen Verständnissen an den Theorien begründet.

In Kapitel 4 werden ein Überblick über die verschiedenen Methoden zur Wissensnutzung und Wissensidentifikation vorgestellt. Die vorgestellte Verfahren in Data Mining sind nur kurz skizziert wurden. Hierzu sei auf die einschlägigen Literaturquellen verweisen.

In Kapitel 5 werden eine Abgrenzung zwischen den Methoden vorgenommen. Die Unterschiede zwischen verschiedenen Methoden in unterschiedlichen Umgebungen wird dargestellt.

In Kapitel 6 werden einige möglich praxisorientierte Anwendungsfelder, auf denen die vorgestellte Konzept später zum Einsatz kommen können, skizziert aufgelistet.

2 Wissensstrukturen und Informationsmanagement

In diesem Kapitel wird die Unterschiede zwischen Daten, Information und Wissen zuerst erläutert. Anschließend wird ein Überblick über die Wissensstrukturen im Zusammenhang mit dem Informationsmanagement angegeben.

Unternehmen müssen immer schneller und flexibler auf Veränderungen am Markt reagieren, um mit der Konkurrenz Schritt halten zu können. Wissen wird dabei zunehmend als eine wettbewerbsentscheidende Ressource angesehen. Es ist zu einem der wichtigsten Werte für Unternehmen geworden. Die Bemühungen der Unternehmen um Generierung und Identifizierung von Wissensstrukturen werden ständig intensiviert. Alle Anstrengungen, Wissen aus externen Quellen in die Unternehmung zu integrieren bzw. selbst neues Wissen aufzufinden bleiben allein jedoch wenig sinnvoll. In vielen Unternehmen, insbesondere in größeren, ist die mangelnde Transparenz über interne Wissensbestände und Experten ein Problem. Im Unternehmen werden vorhandene Wissensbestände oft nicht genutzt, oder den verantwortlichen Führungskräften nicht bekannt sind. Im Rahmen der vorliegenden Arbeit werden also Konzepte eingeführt, Wissen in einer organisierten Art und Weise zugänglich zu machen und schnell zu entdecken. Eine Vorgehensweise, die die effektive Identifikation des im Unternehmen vorhandenen Wissens erhöht, ist die unternehmensinterner Verzeichnisse von Wissensbeständen (anwendungsorientiert strukturierte Informationen). Bei der Repräsentation von Daten, Informationen und Wissen in Computersystemen findet man im allgemeinen keine Unterschiede der Darstellung, es werden immer Zeichen gespeichert und verarbeitet. Erst das In-Beziehung-Setzen der Zeichen durch Syntax, Semantik und Pragmatik führt zur Unterscheidung. Macht man sich die Unterscheidung von Daten, Informationen und Wissen klar, so wird deutlich, dass das Management von Wissen weit über das herkömmliche Management von Information hinausgehen muss.

Die Unterscheidung von Daten, Information und Wissen wird im folgendes erläutert:

Daten sind nicht interpretierte Zeichen und Symbole. Daten haben keine immanente Bedeutung.(z.B. die Zeichenfolge in einer Tabelle: „1 Euro“). Ziel ist es, aus den Daten relevante Informationen zu gewinnen oder – anders ausgedrückt – Daten zu Informationen zu verdichten.

Informationen sind Daten, die durch die Interpretation der Zeichen mit einer Bedeutung versehen sind. (z.B. Größen in einer Tabelle: Der Wert, den man einem Euro

beimisst.). Informationen könnte man definieren als "Kenntnisse über Sachverhalte, die ein Handelnder benötigt, um eine Entscheidung darüber zu fällen, wie er sein Ziel am günstigsten erreichen kann [For99]".
Durch *Wissen* wird Daten eine Bedeutung zugeordnet und somit aus Daten Informationen erzeugt. Hinzukommt die aus dem Wissen gemeinsame resultierende Handlungsfähigkeit: auf der Basis der Informationen können Schlussfolgerungen gezogen werden und sich eigene Aktivitäten anschließen. Wissen kann auf mehreren Ebenen angesiedelt sein: individuelle Personen, organisatorische Einheiten (z.B. Teams oder Abteilungen), Unternehmen, Branchen und Länder. Wissen umfasst also die für wahr und richtig gehaltenen Einsichten, Erfahrungen und Vorgehensweisen einer Person (z.B. Für ein Euro kann man . . .kaufen.) und von Gruppen.
Wissen ist ein entscheidendes Element, mit dessen Hilfe kooperatives Arbeiten koordiniert wird. Unter den Bedingungen der erfolgreichen Konzepte(z.B. Wissensmanagement) repräsentiert Wissen nicht nur individuelles, sondern gemeinsames Verständnis von Zielen und Lösungsmöglichkeiten. Man unterscheidet die folgenden Formen von Wissen:

- Implizites versus explizites Wissen:

Implizites Wissen ist persönliches, verborgenes Wissen und besteht aus Können, Handlungsroutinen, Überzeugungen, Glaubenssätzen und geistigen Schemata. Explizites Wissen ist formales, kodifiziertes Wissen, das leicht durch Worte, Zahlen oder technische Angabe darstellbar ist. Explizites Wissen kann leicht vermehrt werden - anderes ausgedrückt – das dargelegte Wissen kann von den Mitarbeitern der Organisation erlernt, angewendet und erweitert werden, implizites dagegen nicht.

- Know-how:

Know-how interpretieren als " den für die Leistungserstellung relevanten Teil des Wissens" und unterscheiden hierbei:
- *Humankapital*: alle individuellen Kenntnisse und Fähigkeiten der Mitarbeiter,

also personengebundenes Wissen.

- *strukturelles Kapital*: das in einem Betrieb organisatorisch verankerte Wissen, das auch erhalten bleibt, wenn alle Mitarbeiter auf einmal das Unternehmen verlassen würden.

Wissensstrukturen sind Repräsentationen der Organisation von Ideen u.a. in unserem semantischen Gedächtnis. Wissensstrukturen setzt sich also zusammen aus den Verbindungen der einzelnen Ideen, Objekte und Ereignisse untereinander; ist ein Netzwerk von miteinander verknüpften Konzepten im semantischen Gedächtnis.

Die Abbildung von geeigneten Wissensstrukturen liefert die Grundlage für ein effizientes Suchen und Navigieren in dem Bestand an Informationsobjekten. Grundsätzlich ist es notwendig, in einem Wissensmanagmentsystem die zwischen Informationsobjekten bestehenden Relationen kenntlich zu machen. Auf diese Art vermitteln die Beziehungen eines Informationsobjekts zu anderen Beziehungen unterstützend den Kontext des Informationsobjekts, da es seinen Ort im semantischen Gefüge aufzeigt.

Primär werden die in Unternehmen verfügbare Informationsobjekte systematisch nach den hierarchischen Relationen geordnet. Dabei ist zu beachten, dass die Informationsobjekte polyhierarchisch strukturiert werden können und deshalb ein Informationsobjekt mehrere über- bzw. untergeordnete Informationsobjekte haben kann. Da in der hierarchischen Anordnung immer nur eine Baumstruktur zusammenhängend dargestellt werden kann, sind alternative Strukturierungsmöglichkeiten durch Verweisungen auf andere Baumstrukturen anzugeben.

Unter "Informationsmanagement" werden sowohl strategische Aspekte der Informationsversorgung und Informationsverwendung in Unternehmen und Institutionen als auch der Einsatz von Informationssystemen und die Einrichtung der erforderlichen technischen Infrastruktur verstanden. Anders ausgedrückt: Informationsmanagement zu betreiben, heißt dafür zu sorgen, dass die richtige Information zur richtigen Zeit am richtigen Ort ist. Aufgrund dessen ist Information ein grundlegendes Mittel zum Informationsmanagement und für sachbezogene Entscheidungsfindungen. Das Informationsmanagement schließt Kenntnis über den Gebrauch und die Bedeutung von Informationen ein und legt die Wissensbasis der Organisation fest. Das generelle Ziel des Informationsmanagements ist die strategische und operative Institutionalisierung aller mit Informationen und Kommunikation zusammenhängenden Aufgaben und Kompetenzen[Sch91]. So dient Informationsmanagement vor

allem zur Verwaltung und Bereitstellung von Informationen und deren Austausch zwischen Mitarbeitern. Deshalb liefert Informationsmanagement wenige aussagekräftige Informationen, die zur Identifikation von Wissensstrukturen wesentlich beitragen können.

3 Historische Ansätze

In diesem Kapitel wird ein kurzer Einblick über Jean Piagets Theorie über die kognitive Entwicklung des Kinds angegeben. Anschließend wird die Churchman's Notation über die Konstruktion von Erkenntnissystem erläutert.

3.1 Piaget's Theorie

Jean Piaget wurde am 9. August 1896 in der Schweiz geboren. Er war einer der einflussreichsten Denker, der den bedeutenden Beitrag zum Problem der intellektuellen Entwicklung geleistet hat. Er hat Kinder jeden Alters nahezu 50 Jahre lang beobachtet, interviewt und getestet. Die Ergebnisse aus den empirischen Daten sind die Basis seiner Theorie. Seine Theorie über die kognitive Entwicklung des Kinds wirken auf Forschung, und viele andere Gebiete der heutigen Psychologie , Erziehung und Wissenschaft. Nach Piagets Theorie gibt es in der intellektuellen Entwicklung vier Stufen:[Pia91] [Mar75] [Bug93].

1. Die sensomotorische(das Säuglingsalter bis 2 Jahre)
2. Die präoperative(2 bis 7 Jahre)
3. Die der konkreten Operationen(7 bis 11 Jahre)
4. Die der formalen Operationen(11 Jahre und älter)

- Sensomotorische Periode:

Für die Untersuchung des Säuglingsalters hatte Piaget eine Methodologie verwendet, die naturalistisch und informell-experimentell war. Er verließ sich ohne Instrumente (statistischen Verfahren) allein auf genaues Beobachten unter natürlichen Bedingungen , um möglichst viel über den natürlichen Zustand des kindlichen Verhaltens zu erfahren. Um die Mängel des naturalistischen Verfahren auszugleichen, führte er informelle Experimente durch. Er griff in den natürlichen Ablauf der Ereignisse ein, indem das Kind nicht mit einer bestimmten Art von Hindernissen fertig werden konnte. Diese Periode ist im wesentlichen dadurch gekennzeichnet, dass das Kind "eine erste kognitive Orientierung, eine kognitive Umwelt mit den konkreten Dingen im äußeren realen Anschauungsraum aufbaut.". Nach Piaget kann sie selbst in sechs weitere Phasen unterteilt werden:

Im *Stadium 1* (1 Monate) ist der Säugling weitgehend auf seine Reflexe(wie Saugen) angewiesen. Aber der Säugling benützt seine Erfahrung inzwischen auch dazu, die Reflexschemata aktiv zu verändern. Zum Beispiel lernt er, die Brustwarze zu erkennen und nach ihr zu suchen. Angeborene Reflexe und Instinktkoordination sind die Bausteine für die nachfolgenden kognitiven Entwicklung.
Im *Stadium 2* (1 bis 4 Monate) zeigt der Säugling Verhaltensmuster , dass er ein neues Verhaltensmuster oder Schema inzwischen gebildet hat. Nämlich die Hand zum Munde zu führen , um daran zu saugen, indem er zwei getrennte Aktionen- mit der hand zappeln und saugen- koordiniert hat. In dieser Phase hat der Säugling noch keinen ausgereiften Objektsbegriff, um das Objekt zu identifizieren.
Im *Stadium 3* (4 bis 8 Monate) wendet sich der Säugling von Bewegung weg zu Reaktionen auf einen Reiz der äußeren Umgebung. Er bildet eine Zirkularreaktion aus. Beispielsweise stößt er vielleicht an eine Rassel, und wenn er das Geräusch hört, dreht er sich herum und experimentiert, bis er noch einmal dagegen stößt.
Im *Stadium 4* (8 bis 12 Monate) ist es durch das Erscheinen zielgerichteten Verhaltens charakterisiert. Sein Verhalten ist zielstrebig geworden und er lernt auch durch die Interaktion mit seiner Umgebung , welche Beziehung zwischen den Gegenständen bestehen.
Im *Stadium 5* (12 bis 18 Monate) bemüht sich der Säugling aktiv, neue Verhaltensweisen und neue Ereignisse hervorzubringen. Er beginnen mit Variation seiner Bewegung zu experimentieren, um zu sehen, wie sich die Resultate unterscheiden. Er lernt Hilfsmittel zu gebrauchen, um ein Ziel zu erreichen. Er handelt intelligent.
Stadium 6 (18 bis 24 Monate)bildet den Übergang zum symbolischen Denken. Die Entwicklung stellt im weitesten Sinne einen Prozess der Dezentrierung dar. Das heißt, der Säugling ist völlig auf sich selbst zentriert. Er unterscheidet weder zwischen Selbst und Umgebung noch zwischen Wunsch und Realität. Er weiß noch nicht , ob die Gegenständen voneinander abhängig oder unabhängig sind ? So erfährt der Säugling die Welt als eine Folge wechselnder und unverbundener Bilder

- Präoperationale Periode:

In dieser Periode, die bereits durch das sechste Stadium der sensomotorischen Intelligenz eingeleitet wird, wird die Fähigkeit entwickelt, nicht nur mit Dingen selbst, sondern auch mit ihren verinnerlichten Repräsentationen zu agieren. D.h. das

Kind erkennt das Symbol (Bild) eines Objektes und ist in der Lage, zwischen Objekt und dem Symbol des Objekts zu unterscheiden und dennoch beide aufeinander zu beziehen. Zu diesen symbolischen Repräsentationen gehört auch die Sprache.
Gegenüber der späteren Phase des konkret operationalen Denkens, unterscheidet sich das Kind der voroperationalen Stufe durch einen noch sehr hohen Grad an Konkretheit im Denken und einer starken Verbindung des Denkens mit dem Handeln. Ein weiteres Unterscheidungsmerkmal wäre die Zentrierung, d.h. die Konzentration des Kindes auf bestimmte hervorstechende Merkmale eines Objekts oder Sachverhalts.
Dies demonstrierte Piaget mit seinen Umschüttversuchen:
In zwei gleichartigen Gefäßen befinden sich zwei gleiche Flüssigkeitsmengen. Auf die Frage in welchem Gefäß mehr enthalten ist, antwortet das Kind, dass in beiden die selbe Menge wäre. Gießt man nun vor den Augen des Kindes eine der beiden Flüssigkeiten in ein schmäleres Gefäß, so wird die Antwort des Kindes lauten, dass sich dort eine größere Flüssigkeitsmenge befindet, da es das hervorstechende Merkmal der Höhe der Flüssigkeitssäule mit der Menge gleichsetzt. Das Kind realisiert noch nicht die Erhaltung des Ganzen.

- Periode der konkreten Operationen:

Unter der konkreten Operationen versteht man, dass das Kind im Gedanken mit konkreten Objekten oder ihren Vorstellung operieren kann. Piaget führte den Begriff der Operationen in die Entwicklungspsychologie ein. Die wesentlichsten Merkmale dieses Begriffs sind:
In dieser Phase sind die gedanklichen Operationen weiterhin an anschaulich erfahrbare Inhalte gebunden, sie zeichnen sich jedoch durch eine größere Beweglichkeit aus. Verschiedene Aspekte eines Gegenstandes oder Vorgangs können gleichzeitig erfasst werden und zueinander in Beziehung gesetzt werden.
Aktivität: gemeint ist, dass eine Operation eine abstrakte Tätigkeit darstellt.
Systematisierung: Piaget vertritt die Ansicht, dass Operationen nie isoliert Vorkommen, sondern immer in Verbindung mit einem ganzen System von Operationen auftreten.
Dezentrierung: Das Kind lässt sich nicht mehr von einem hervorstechenden Merkmal des Subjekts beeinflussen, sondern ist jetzt in der Lage auch andere, nicht so offensichtliche Merkmale, wahrzunehmen und in sein Denken miteinzubeziehen.

Reversibilität: Reversibilität ist die Fähigkeit, eine und dieselbe Handlung in beiden Durchlaufsrichtungen auszuführen und zwar im Bewusstsein davon, dass es dieselbe Handlung ist.
In dieser Phase sind die gedanklichen Operationen weiterhin an anschaulich erfahrbare Inhalte gebunden. Sie zeichnen sich jedoch durch eine größere Beweglichkeit aus. Verschiedene Aspekte eines Gegenstandes oder Vorgangs können gleichzeitig erfasst werden und zueinander in Beziehung gesetzt werden.

- Periode der formalen Operationen:

Ist das Denken in der Periode der konkreten Operation noch sehr auf konkrete Handlungen und Wahrnehmungen bezogen, so beginnt sich das formal operative Danken von den konkreten Objekten zu lösen und wird dabei immer mehr formalisiert. Das heißt Konkretes wird immer mehr als Spezialfall von Hypothetischem gesehen. Der Übergang von konkret operativem zu formal operativem Denken ist interindividuell sehr verschieden und der Beginn etwa ab Vollendung des 11.Lebensjahres anzu setzen. Außer der Loslösung vom Konkreten hin zum hypothetisch Möglichen, nimmt auch die Integration und Systematisierung zu immer umfassenderen Gesamtstrukturen weiter zu.
Die hypothetisch- deduktive Vorgehensweise wird immer mehr zur konsequenten Strategie, was sich auch darin äußert dass vermehrt "Wenn..., dann", "Was wäre wenn..." und ähnliche Formulierungen auftreten. Das Denken bezieht sich nun nicht mehr auf konkret reale Objekte, als vielmehr auf Aussagen über diese, d.h. die in Aussagen gefassten Ergebnisse konkreter Operationen werden weiteren Operationen unterzogen. Diese heißen Operationen zweiten Grades.
Besonders kennzeichnend für das formaloperatorische Denken ist die Kausalanalyse im Natur- oder Kausalwissenschaftlichen Bereich. Tritt die Frage nach der Ursache eines bestimmten Effekts auf, werden zuerst alle möglichen Wirkfaktoren bestimmt und analysiert, und im weiteren auf ihre Wirkung hin untersucht. Das hypothetische Ergebnis wird anschließend am Experiment verifiziert.

3.2 Churchman's Notation

Nach der Entstehung der Informationstechnologie hat die anwachsende Fähigkeit des Informationssystem (wie CBIS), deren Methodologien stark von traditionell wissenschaftlichen und technischen Paradigmen abhängig sind, und der weitverbreiteten Einsatz in Organisation zwei philosophische Probleme verursacht:
Erstens, Die Informationstechnologie verhindert die Fachkräfte, welche eine zusammenhängend historische Perspektive aus ihrer Entstehung und Folge entwickeln wollte.
Zweitens, Es ist nicht gelungen, eine überzeugend logisch Grundlage für ihrer intellektuellen Rechtfertigung zu schaffen.
Die Problematik, mit der sich Churchman beschäftigt , veranlasst ihn, ein neues System zu konstruieren. Im Jahr 1971 veröffentlicht er das philosophische Konzept — Inquiring Systeme. Die Systeme sind im Grunde sehr komplexe, selbstlernende, und selbstüberprüfende Systeme. So wird eine Klassifikation von verschiedenen Systeme (Leibnizsche, Lockesche, Kantische, Hegelsche und Singersche) entwickelt. Die Verwendung von Systemsphilosophie für die Konstruktion der Erkenntnissystem von Elementen ist empfehlenswert. Er begründet folgendes: [Chu73]

a) Die Konstruktion eines Erkenntnissystem sollt das Konzept der Unterscheidung zwischen Wirklichkeit und Nichtwirklichkeit anpassen. Und nur ein System , das die Informationen zu der „ganzen" Wirklichkeit in Beziehung setzt, kann diese Ziel erreichen. Leibniz' Theorie Monade kann zu der Konstruktion dieses System beitragen, seitdem die Monade die Gesamtwirklichkeit symbolisch in sich enthalten.
b) Lockesche „ Essay on human Understanding" ist bedeutsam für die Konstruktion von Erkenntnissysteme, in denen es den menschlichen Verstand as ein tabuale Rasa , keine angeborenen Ideen, betrachtet. Die Systemkonstruktion, die auf diese Prinzip basiert ist, kann jede beliebige „Information" aufnehmen, genauso wie Eingabegrößen aus jeder anderen Quelle behandeln , und eine klare Wahrnehmung des Verständnis entwickeln. Genau bei dem menschlichen Verstand ist es schwierig zu konstruieren.
c) In ein Erkenntnissystem müssen es einige angeborene Bearbeitungsfähigkeiten ohne Berücksichtigung von Eingabe und Ausgabe vorhanden sein. Kantische

„Critique of Pure Reason “ und die Konzeptualisierung der a priori Ideen können für die Konstruktion hilfreich sein.
d) Hegelscher Ansatz kann bei der Konstruktion von Erkenntnissystem verwendet werden, seitdem er die Objektivität in verständlichen und einen selbstlernenden Prozess führt, der auf eine dialektischen Methode basiert .
e) Singersche Ideen, die einen metrologischen Ansatz(Messtheorie) zum Konstruieren betonen, können bei der Konstruktion hilfreich sein, seit die formale Messung, die Maßeinheit, und die Standardverfahren extrem wichtig für die diese Systeme sind.
Churchman hat bei der Konstruktion von Erkenntnissysteme drei Grundmodelle von Erkenntnissystem eingeführt, die dem Erkenntnissystem zur Verfügung stehen:

- Das Demokritische Modell: der Mechanismus

Eine anziehende Gesichtpunkt des Demokritischen Modells ist seine Abstraktionskraft. Denn er die Veränderung als voraussagbare Veränderung bestimmter Elemente in Wirklichkeit beschreibt. Mit ein paar einfachen mathematischen Gleichung kann man weite Bereiche einfangen. Aus heutigen Sicht wird das Wissen mithilfe der mathematische Verfahren aus den historischen Datenbeständen abgeleitet. Zum Beispiel, ein Demokritisches Erkenntnissystem würde vielleicht versuchen, aus den täglichen Ereignissen an der Wertpapierbörse gewisse Grundinformationen(tägliche Umsatzvolumen in verschiedenen Sektoren) abzuleiten, die die Änderungen erklären und Voraussagen für den gesamten Markt oder Teilmarkt erlauben.

- Das Aristorelische Modell: Teleologie

In diesem Modell werden die Elemente der Natur(Objekte) als zweckgerichtet aufgefasst, d.h. wegen ihren Verhaltensweise ihnen bestimmte Ziele zugeschrieben. Nach diesem Modell hat jedes Element verschiedene Wahlmöglichkeiten, die ihm zukommenden Ziele zu verfolgen. Mit dem teleologische Modell scheint ein einfache Beschreibung und Erklärung des Verhaltens von Individuen zu ermöglichen. Wir benutzt diese Modell, um unsere Informationen darzustellen, und die Ereignisse in der beobachtbaren Umgebung zu erklären. Ein Aristotelische Erkenntnissystem würde die Wertpapierbörse unter dem Gesichtspunkt des zweckgerichteten Verhaltens von Marktteilnehmer betrachten wollen. Es würde jedem Anleger

bestimmte Ziele unterstellen und mit Hilfe des teleologischen Modells das Gesamtverhalten des Markts als zielsuchendes Verhalten zu erklären versuchen.(Kauf- oder Verkaufsignale).

- Das Karneadische Modell: Wahrscheinlichkeit

Ein Grundbegriff im Karneadischen Modell ist die Wahrscheinlichkeit, die messbar sind. Mit Hilfe des Modells ist das Erkenntnissystem in der Lage, nach statistischen Strukturen bei den Ereignissen zu suchen. Das Modell ist aus heutigen Sicht in vielen Bereiche auch sehr fruchtbar, in denen die Menge der Ereignisse und Informationen jede Demokritische oder Aristotelische Deutung auszuschließen scheint. Nach dem karneadischen Modell verhält sich der Wertpapiermarkt ganz wie ein ziemlich komplizierter Zufallsaparat.
Die oben eingeführte Modelle lassen sich auch miteinander erweitern. Beispielsweise kann man das karneadische Modell erweitern, um zu Demokritschen Modell zu kommen, indem man passende Wahrscheinlichkeitsverteilungen einführt, die den Ereignissen eine Gewichtung 0 oder 1 zuschreiben.
Nach diesen Modelle konstruierte System schafft es, Wissen besser zu generieren und schnell zu entdecken. Da das Erkennenssystem besser aufgebaut sind, das die logische Beziehung zwischen Informationsobjekten ermitteln und interpretieren kann.

4 Methoden zur Identifikation von Wissensstrukturen

In dieser Kapitel werden einige Methoden und Techniken vorgestellt, die sich in Verschiedenen Umgebungen zum Analysieren und zur Identifikation von Wissensstrukturen verwenden lassen.

4.1 Knowledge Discovery in Database und Data Mining

Mit der Entwicklung der Informationstechnologie haben sich in den vergangenen Jahren die Möglichkeiten, alle anfallenden Informationen wie Marktkennzahlen, Kundendaten zu sammeln, zu archivieren, entscheidend verbessert. Durch die weitverbreitete Anwendung von Barcodes und Scanner-Technologie bei den meisten kommerziellen Produkten sowie durch die zunehmende Computerisierung vieler Unternehmensbereiche, beispielsweise durch die Bezahlung mit Kredit- oder Kundenkarten, haben sich die Datenmengen sprunghaft erhöht. Die neue Kommunikationstechnologie ermöglichen den großen Unternehmen, die Daten in verschiedenen Orts zuzugreifen, zu verwalten. Was dazu führt, dass die Datenbanken bislang in bekannte Dimensionen wachsen. Das Phänomen der Datenflut führt dazu, dass das Informationsangebot für das Management einerseits zunimmt und andererseits die Versorgung des Management mit relevanten Daten zunehmend schwieriger wird
Es ist klar, dass diese Datenbestände mit den traditionellen Analysetechniken wie SQL-Abfrage oder Bereichten nicht mehr bewältigt werden können. Aus alle diesen Gründen wächst das Bedürfnis nach neue Konzept und automatischen Auswertungsmechanismen, die nützliches Wissen aus großen Datenbanken herausfiltern können. In den folgenden Abschnitten soll ein Einblick über KDD und Data Mining gegeben werden, welche Methoden und Techniken entwickelt werden, um neues Wissen zu entdecken.

4.1.1 Begriffsdefinitionen

Die Sichtung verschiedener Definitionen der Begriffe KDD und Data Mining führt auf eine Vielzahl verschiedener Begriffsfindung. Zum Beispiel “Data Mining is the process of discovering advantageous patterns in data (John 1997) “ oder “Data Mining is a decision support process where we look in large databases for unknown and unexpected patterns of information(Parsaye 1996)”[Säu00]. In vielen Fälle

werden die Begriffe Data Mining und KDD synonym verwendet. Data Mining wurde vorwiegend von Statistikern, Datenanalysten und Forscher von Datenbank-managementsystem, KDD von den Vertretern für künstliche Intelligenz und das maschinelle Lernen verwendet[Fay96a]. In den letzten Jahren hat sich in der wissen-schaftlichen Literatur ein auf bereite Basis akzeptiertes Verständnis der Begriffe Data Mining und KDD durchgesetzt. Fayyad hat 1996 eine anerkannte Definition eingeführt:

> *Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*

Mit „no-trival“ wird gefordert, dass ein Such- oder Schlussfolgerungs-algorithmus zur Anwendung kommt, um Data Mining von einfache statistischen Auswertung unterscheiden zu können. Der Begriff „valid“ besagt, dass die Gültigkeit der Daten Muster über die verwendeten Daten hinaus überprüft werden muss. Mit anderen Worten, die Gültigkeit der in einer Stichprobe gefundenen Muster in anderen Stichproben zu überprüfen. Die Forderung nach neuen(novel), potentiell nützlichen(potentially useful) und verständlich Mustern (understandable patterns) sind pragmatischer Natur und unmittelbar verständlich. Dieser in der Literatur auch als „Knowledge Extraction“ oder „Data Analysis“ bezeichneter Prozess ist darauf ausgerichtet, in umfangreichen Datenbeständen implizit vorhandenes Wissen zu entdecken und explizit zu machen[Düs98]. So ist es beispielsweise im Wal-Mart durch den Prozess des Knowledge Discovery in Databases möglich, auf der Grund-lage von Transaktionsdaten(ca. 20 Millionen täglich) komplementäre Beziehungen zwischen einzelnen Artikeln zu identifizieren. Komplementäre Artikelbeziehung sind Ausdruck des Einkaufsverhaltnis von Kunden, wie die Frage „ wenn Kunden Clips und Bier kaufen, dann kaufen sie auch häufig Salzstangen“. Die daraus ermittelte Zusammenhänge ermöglichen die Festlegung von Maßnahmen, die eine Erhöhung der Kundbindung erzielen und sich z.B. in der räumlichen Platzierung von identi-fizierten Nachfrageketten(Chip, Bier und Salzstange) niederschlagen.

Bei der Begriff Data Mining dachte man dabei hauptsächlich an Algorithmen und Computerprogramme, mit denen die Beziehung zwischen den betrachteten Daten, die Daten Muster, ermittelt werden konnten. Entsprechend kann definiert werden:

> *Data Mining ist die Anwendung spezifisches Algorithmen zur Extraktion von Mustern aus Daten*[vgl. Fay96a]

Auf der Basis der angeführten Definitionen soll Data Mining hier generisch und prozeßorientiert definiert werden. So wird Data Mining als integrierter Prozess verstanden, der durch die Anwendung von Methoden auf einen Datenbestand Muster identifiziert[Ben99]. Die gefundenen Muster müssen für einen möglichst groß Teil der Daten Geltung haben und bislang unbekannte, potentiell nützlicher und leicht verständliche Zusammenhänge in den Daten zum Ausdruck bringen. Aus den ermittelten Beziehungen wird schließlich durch Interpretation und Evaluation explizites Wissen abgeleitet[Düs98]. Nach der Definition wird der KDD-Prozeß im folgende vorgestellt.

4.1.2 Der KDD-Prozeß

Der KDD-Prozß ist ein komplexer Prozess, in dem mehrere Phasen iterativ durchlaufen werden und in den Mensch und Maschine interaktiv ihre jeweiligen Stärken einbringen[vgl. Fay96a]. Erst die gezielte Abfolge von Prozessschritten und auf das Problem angepassten Verfahren ermöglichen die Entdeckung von neuartigem nützlichem und nachvollziehbarem Wissen in Datenbanken. In Abbildung 4.2 ist der KDD-Prozess mit den zugehörigen Schritte dargestellt. Die Teilschritte des in Abbildung dargestellten Modells werden im folgenden kurz erläutert:

- Selection:

Zunächst wird eine Teilmenge der Roh-Daten ausgewählt, die als Basis für das weiter Vorgehen dient. Dies kann notwendig sein, wenn nur ein bestimmter Teil bereich der Daten untersucht werden soll, z.B. wenn in einer Kundendatenbank nur ein bestimmter Alterinterval von Interesse ist. In dieser Phase wird geprüft, welche Daten notwendig und verfügbar sind, um das gesetzte Ziel zu erreichen.

- Preprocessing:

Diese Phase dient der Säuberung und Aufbereitung der angewählten Daten. Die Säuberung enthält daher Maßnahmen, um bspw. fehlende Werte durch Abgleich mit anderen Datenquellen oder durch entsprechende Default-Werte zu ergänzen oder um mittels übergreifender Interitätsregeln Datenfehler zu erkennen und zu erheben.
Die Erkenntnisse, die der Anwender in dieser Phase der Vorarbeitung über den Datenbestand gewinnt, kann Hinweise auf die Verbesserung der Datenqualität des operativen Systems geben.

- Transformation:

In diesem Schritt werden die Daten für die nachfolgende Analyse vorbereitet. Beispielsweise mehrere Bestellungen eines Kunden innerhalb eines Monats zu einem Datensatz zusammengefasst werden, der die Anzahl der Bestellungen und deren durchschnittlichen Wert enthält. Dieser Schritt kann schon notwendig sein, um die Datenmenge einzuschränken, die zur Verarbeitung ansteht. Durch Reduktion von Dimensionen der Daten werden die zu betrachten Variablen reduziert.

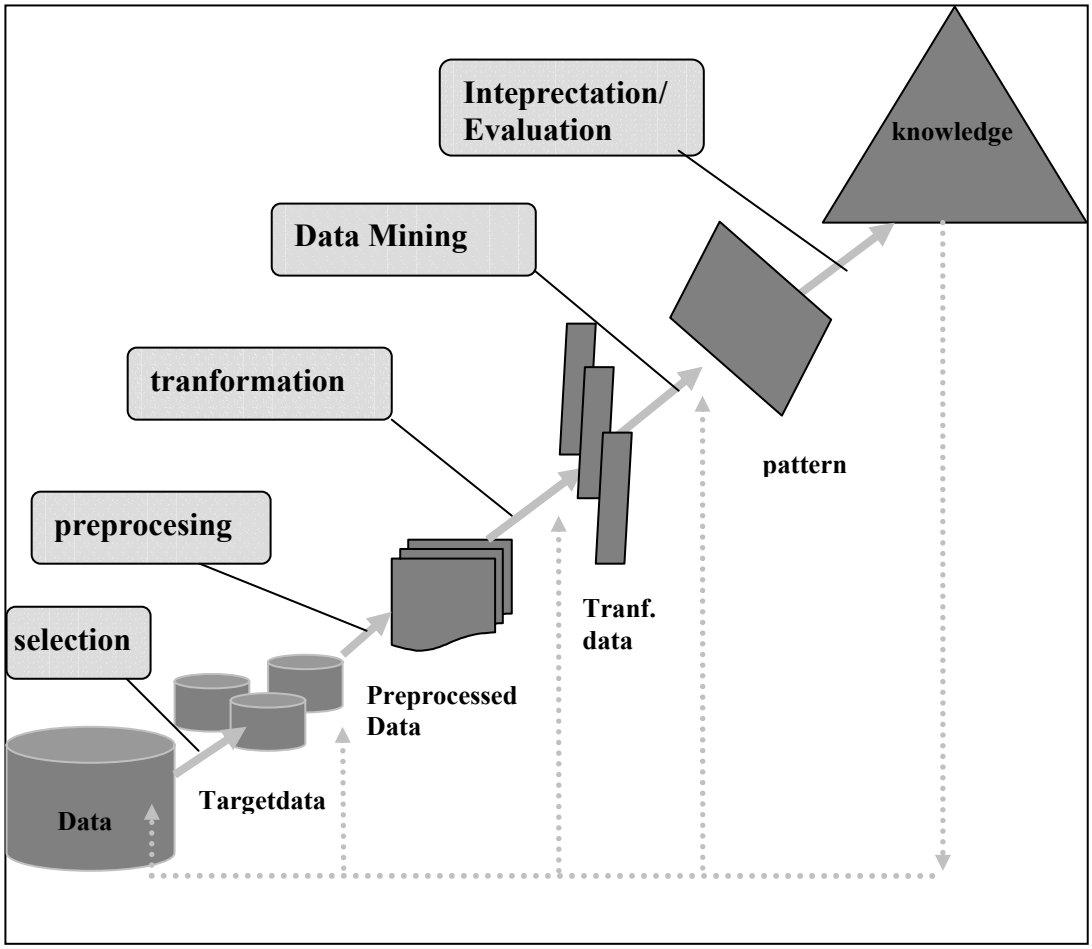


Abb.4.1: Der KDD-Prozeß [Fay 96a]

- Data Mining :

Die Data Mining Phase stellt den wichtigsten Bestandteil des KDD-Prozesses dar. Durch Methodenauswahl(Clusteranalyse, Neuronale Netze etc.) werden interessierende Muster und Beziehungen in den Daten erkannt. Im allgemeinen wird dieser Schritt iterativ und interaktiv mit den Anwender durchgeführt. Eine detaillierte Darstellung der eingesetzten Verfahren folgt im nächsten Abschnitt.

- Interpretation/Evaluation: Hier erfolgt die Interpretation und Bewertung der entdeckten Muster. Eine Evaluation der Resultate erfolgt meist in Zusammenarbeit

mit den einzelne Fachbereichen. Wahlweise erfolgt in dieser Phase eine Rücksprung in eine der vorherigen Phasen. So ist die Auswahl einer anderen Data Mining Technik kann wünschenswert sein.
Am Ende des Prozesses soll das entdeckte Wissen soweit wie möglich in den Arbeitsablauf integriert werden, Da das entdeckte Wissen in einer dynamischen Welt recht schnell veraltet und dann u.U. zu Fehlentscheidungen führt. Auch ein Abgleich im Hinblick auf Widersprüche mit bisherigem oder erwartetem Wissen sollte vorgenommen werden.

4.1.3 Data Mining-Methoden

Nachdem im letzen Abschnitt die Phasen des KDD-Prozesses einer genaueren Betrachtung unterzogen wurden, sollen in diesem Abschnitt zunächst die Teilziele vorgestellt werden, die mit Data Mining verfolgt werden können. Daran anschließend soll anhand von vier grundlegenden Operationen ein Einblick in die Arbeitsweise eines Data Mining-Systems gegeben werden.

4.1.3.1 Aufgaben von Data Mining

Wie schon in der Beschreibung des KDD-Prozesses zu sehen, liegen die Aufgaben eines Data Mining-Systems in dem weitgehend automatischen Auffinden von bisher unbekannten Zusammenhängen in großen Datenmengen. Darauf aufbauend könne zwei Teilziele definiert werden: Vorhersage und Beschreibung [Fay96a].
Bei der Vorhersage werden ausgehend von den Rohdaten neue Muster oder Regeln extrahiert. Beispielsweise wird in Kundendaten einer Bank die Kreditwürdigkeit eines Kunden als zu vorhersagende Variable definiert. Diese Kreditwürdigkeit könnte in Form einer logischen Regel aus den anderen Kundenattributen erklärt werden: „Kreditwürdigkeit ist hoch, wenn Haushaltseinkommen größer als x DM pro Monat und keine Kontosperre im letzten Quartal und ...“. Aufgrund dieser Regeln kann die Kreditwürdigkeit von neuen Kunden anhand ihrer persönlichen Merkmale eingeschätzt werden.
Bei der Beschreibung werden eher verständliche Muster oder Abhängigkeiten in den Daten gesucht. Ein typisches Beispiel hierfür ist eine Warenkorbanalyse, wobei Gruppen von Produkten gefunden werden, die häufig gemeinsam gekauft werden. Diese Beziehungen zwischen den einzelnen Produkten können durch Wenn/Dann-

Regeln beschrieben werden. Die Unterscheidung zwischen beschreibenden und vorhersagenden Modellen ist allerdings nicht sehr ausgeprägt, da vorhersagende Modelle einerseits einen beschreibenden Charakter haben, andererseits beschreibende Modelle ebenfalls zu Vorhersage genutzt werden können.

4.1.3.2 Überblick über die Data Mining-Methoden

Im folgenden sollen vier grundlegende Operationen vorgestellt werden, die mit einem Data Mining-System erreicht werden können (siehe [Fay96a], [Kra98], [Nak98]) . Damit soll ein Einblick in die Arbeitsweise eines solchen Systems gegeben werden. Auch eine ungefähre Abschätzung der Art der produzierten Resultate ist damit möglich.

- Segmentierung:

Die Segmentierung zielt auf die Aufspaltung der Daten in interessante und sinnvolle Teilmengen oder Klassen, sogenannte Segmente. Auf der Basis von Distanzmaßen soll dabei innerhalb eines Segments eine höchstmögliche Homogenität, zwischen den Segmenten eine größtmögliche Heterogenität erreicht werden. Der Grad der Homogenität und die Anzahl der Segmente kann vom Benutzer über Parameter bestimmt werden. Die Segmentierung wird häufig zur Einteilung von Kunden in Zielgruppen verwendet, um eine möglichst zielgruppenorientierte Marketing-Aktivität zu realisieren. Dazu werden die Kundendaten und Transaktionen über einen gewissen Zeitraum analysiert, wobei ähnliche Verhaltensmuster identifiziert werden.

- *Klassifikation:*

Bei dieser sehr häufig verwendeten Methode wird ein Datensatz in eine vordefinierte Klasse oder Gruppe eingeordnet. Dabei werden historische Daten verwendet, um ein Klassifikationsmodell zu entwickeln(ein solches Modell wird auch Klassifikator genannt), das hilft, die bestehenden Daten und deren Verhalten in der Zukunft zu verstehen. Eine typische Anwendung ist beispielsweise die Einordnung von Versicherungsnehmern in Risikoklassen bei einer Autoversicherung.
Der Unterschied zwischen der Klassifikation und der Segmentierung liegt darin, dass bei der Klassifizierung vordefinierte Klassen verwendet werden, während bei der Segmentierung diese erst generiert werden. Aus diesem Grund können mit dieser Methode auch keine Klassen entdeckt werden, die zuvor noch unbekannt oder nicht definiert wurden.

- Abhängigkeitsanalyse:

Mit der Abhängigkeitsanalyse werden Modelle generiert, die Beziehungen und Abhängigkeiten zwischen Variablen beschreiben. Dabei können zwei Ebenen unterschieden werden:

Auf der *strukturellen* Ebene werden die Zusammenhänge der Abhängigkeiten bestimmt, auf der quantitativen Ebene die (numerische) Stärke der Abhängigkeit. Bei einer Warenkorbanalyse kann ein Data Mining-System beispielsweise auf der strukturellen Ebene feststellen, dass zwischen dem Kauf von Brot und dem Kauf von Butter eine Abhängigkeit besteht.

Auf der *quantitativen* Ebene kann die Wichtigkeit der Aussage festgestellt werden, indem die Abhängigkeit beispielsweise mit einer Wahrscheinlichkeit von 80% eintritt. Ebenfalls in den Bereich der Assoziationen fallen sequentielle Assoziationen bzw. sequentielle Muster. Sequentielle Assoziationen setzen Datenhistorien voraus und zielen darauf ab, dass innerhalb eines Zeitraumes Beziehungen zwischen dem Auftreten verschiedener Ereignisse eintreten.

Ein Einsatzgebiet ist das Aufdecken von Betrugsfällen im Kreditkartenbereich. Es wurde festgestellt, dass Kreditkartenbetrug oft nach regelmäßigen Schemata abläuft, was sich in der zeitlichen Abfolge bestimmter Transaktionen ausdrückt.

- Abweichungsanalyse

Die Abweichungsanalyse beschäftigt sich mit Objekten, die sich keinem Muster eindeutig zuordnen lassen. Bei diesen „Ausreißern“ kann es sich um fehlerfreie, interessante Merkmalsausprägungen handeln oder aber auch um fehlerhafte Daten, die keine realen Sachverhalte beschreiben. Die Zielsetzung der Abweichungsanalyse besteht darin, die Ursachen für die untypischen Merkmalsausprägungen des Ausreißers aufzudecken. Wird ein Ausreißer im Datenbestand identifiziert, so werden alle assoziierten Datenbestände durchsucht, um die Einflussfaktoren zu erklären, die zu einer abweichenden Merkmalsausprägung geführt haben. Auch signifikante Änderungen in Bezug auf vorher definierte oder gemessene Werte sollen erkannt werden. Dies wird vor allem im Bereich des Controlling verwendet, beispielsweise für Soll-Ist-Vergleiche. Bei der Zeitreihenanalyse, einer Variante der Abweichungsanalyse werden die Abhängigkeiten in Form von sequentiellen Mustern untersucht, die bezogen auf einen bestimmten Zeitraum ein ähnliches Verhalten aufweisen. Dies ist vor allem für die Bestimmung von Trends nützlich. Beispielsweise kann das Kaufverhalten von Kunden über einen Zeitraum hinweg analysiert werden und Abhängigkeiten zwischen zeitlich verschiedenen Kaufvorgängen ermittelt werden.

In Abb.4.2 werden die Beziehungen zwischen den Zielen von Data Mining und den Methoden verdeutlicht:

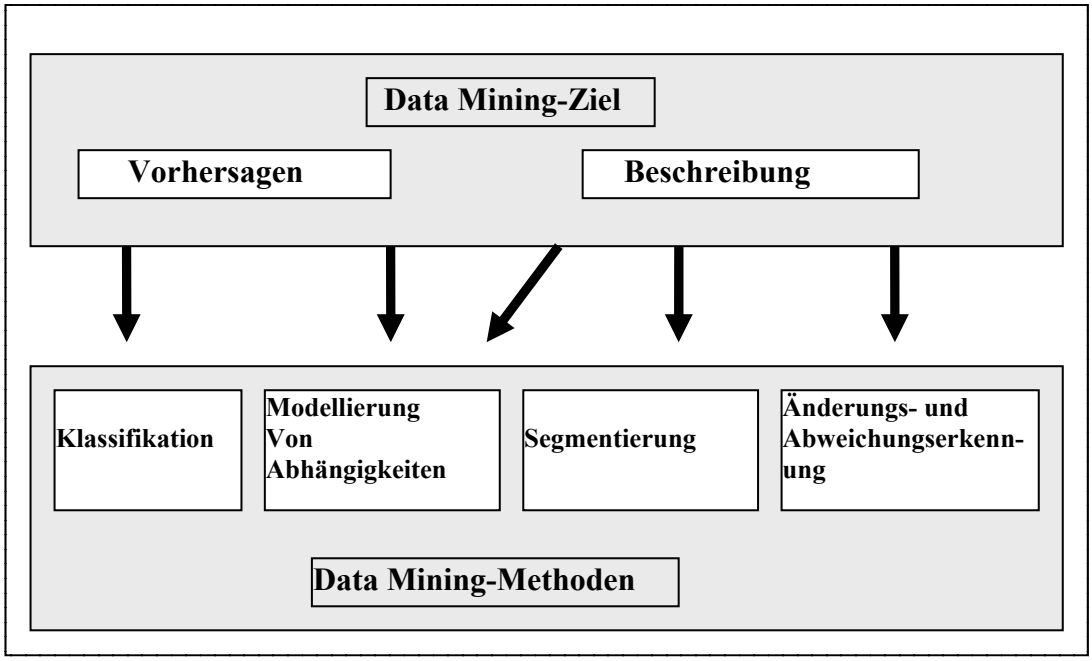


Abb.4.2: Verbindung zwischen Data Mining-Zielen und Data Mining-Methoden

4.1.4 Data Mining-Techniken

In diesem Abschnitt soll ein Überblick über die gängigsten Data Mining-Techniken gegeben werden. In Abbildung 4.3 ist eine Zuordnung dieser Techniken zu den weiter oben vorgestellten Methoden zu sehen.

Einige Techniken können für mehr als eine der grundlegenden Methoden verwendet werden, wobei zur Lösung eines Problems oft einen Kombination von zwei oder mehreren Techniken verwendet wird. Darüber hinaus muss berücksichtigt werden,

dass sich dieses Forschungsgebiet ständig weiterentwickelt, so dass diese Darstellung keinen Anspruch auf Vollständigkeit erhebt. Von den vielen Methode, die für Data Mining verwendet werden können, werden nachfolgend nur die am meisten verwendeten Methoden kurz erläutert.(siehe [Fay98b], [Kra98], [Küp98], [Bac96], [Bor98], [Alp00])

dass sich dieses Forschungsgebiet ständig weiterentwickelt, so dass diese Darstellung keinen Anspruch auf Vollständigkeit erhebt. Von den vielen Methode, die für Data Mining verwendet werden können, werden nachfolgend nur die am meisten

verwendeten Methoden kurz erläutert.(siehe [Fay98b], [Kra98], [Küp98], [Bac96], [Bor98], [Alp00])

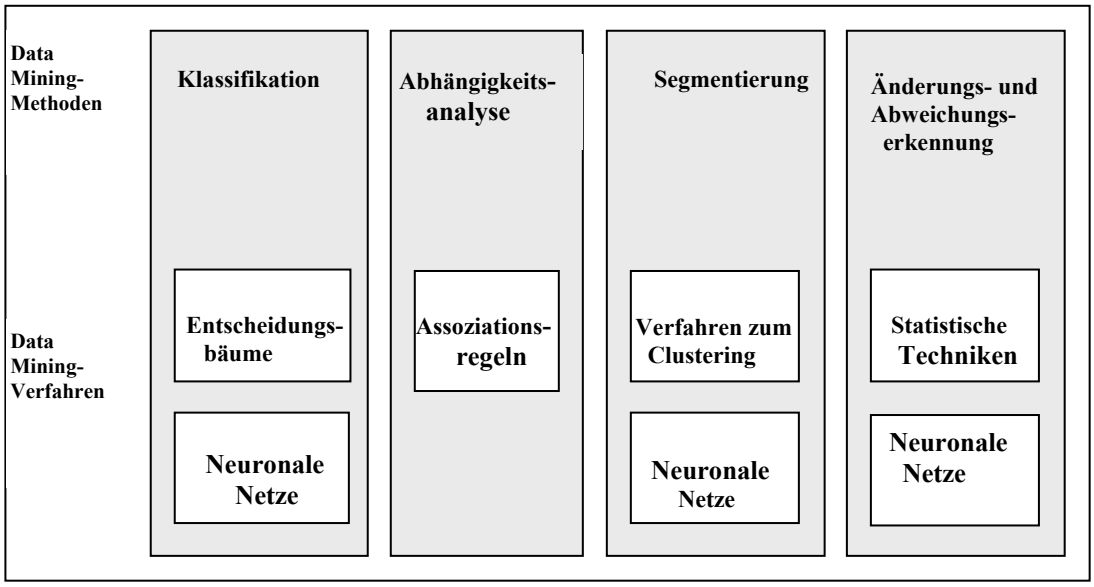


Abb.4.3: Zuordnung von Data Mining-Techniken zu Data Mining-Methoden

4.1.4.1 Entscheidungsbäume

Entscheidungsbäume gehören zu den bei Data Mining-Systemen am häufigsten verwendeten Ansätzen, und werden meist zur Klassifikation verwendet. Die Menge der Datensätze wird dabei in einer baumähnlichen Struktur in Kategorien eingeordnet. Abbildung 4.4 zeigt einen solchen Entscheidungsbaum zur Vorhersage des ROI (Return of Investment). Der markierte Pfad zeigt eine Folge von Entscheidungen, die zu einem besseren Ergebnis führen. Der Algorithmus zum Aufbau eines Entscheidungsbaumes beginnt mit der Suche nach einer sinnvollen Unterteilung des Wurzelknotens. Dazu wird eine Menge von Trainingsdatensätzen Herangezogen, in denen das abhängige Attribut (hier der ROI) bereits bekannt ist. Der Algorithmus sucht nun in den unabhängigen Attributen nach demjenigen, welches die Daten am besten in die gewünschten Kategorien unterteilt. Dieser Prozess wird für jeden entstehenden Baumknoten neu durchgeführt, bis kein passendes Attribut mehr zum Teilen gefunden werden kann. Das Auswahlverfahren zur Bestimmung der Attribute ist für die Generierung des Baumes entscheidend. Häufig implementierte Algorithmen sind sogenannte CARTs (*classification and regression trees*) und CHAIDs *(chi-squared automatic interaction detection*):

- Bei einem CART-Algorithmus wird die Attributauswahl durch Maximierung des

Informationsgehalts gesteuert. Dazu wird zu jedem Attribut ein Schwellwert gesucht, der eine optimale Trennung der Daten in Bezug auf die Klassifikation zulässt. Grundsätzlich lässt sich formulieren: Je höher der Informationsgehalt eines Attributs

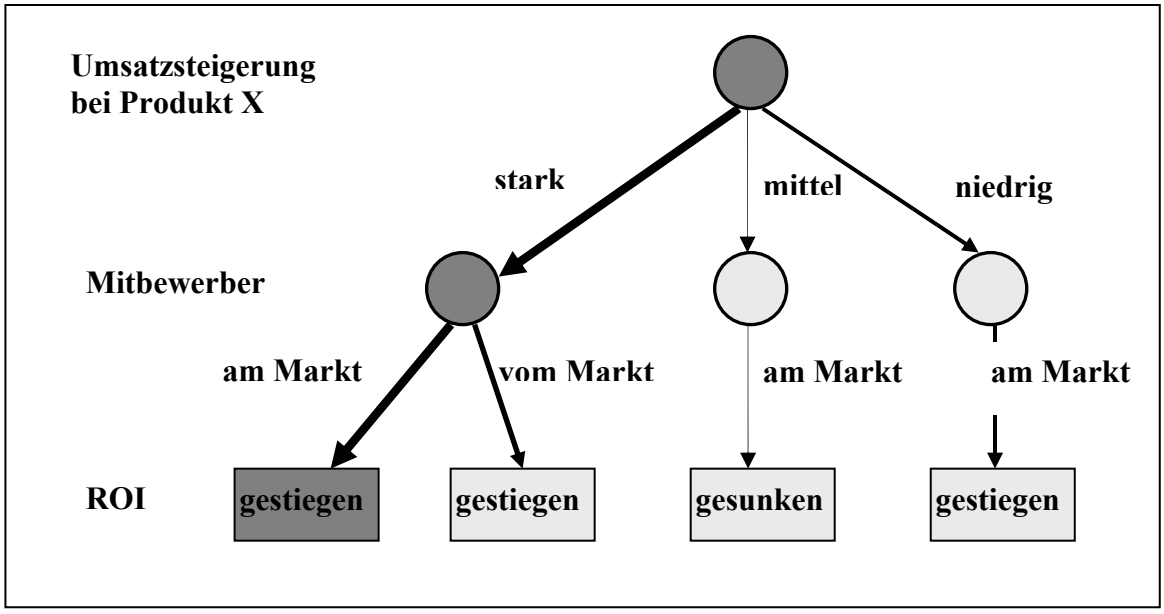


Abb.4.4: Beispiel eines Entscheidungsbaumes zur Vorhersage des ROI

in Bezug auf die Zielgröße, desto weiter oben im Entscheidungsbaum findet sich dieses. Eine Eigenart der CART-Bäume ist die Tatsache, dass durch die Teilung der Attribute durch einen festen Schwellwert nur Binärbäume erzeugt werden können. Binärbäume sind dadurch gekennzeichnet, dass an jeder Verzweigung genau zwei Äste vorhanden sind.

- Bei CHAID-Algorithmen wird zur Attributauswahl der Chi-Quadrat-Unabhängigkeitstest verwendet. Dieser wird benutzt, um eine Aussage über die Abhängigkeit zweier Variablen zu treffen. Es wird dazu eine Kennzahl, der sogenannte „Chi-Quadrat-Abstand“, berechnet. Bei der Attributauswahl wird dann die Variable mit dem größten Abstand zur Zielgröße berücksichtigt. Im Gegensatz zum CART-Algorithmus können hier an jedem inneren Knoten des Baumes auch mehr als zwei Verzweigungen vorgenommen werden, wenn dadurch die Trennqualität erhöht wird. Die dadurch entstehenden Mehrfachbäume haben eine kompaktere Form der Darstellung und führen zumindest theoretisch zu besseren Ergebnissen. Allerdings ergeben sich bei der Verwendung von numerischen Daten bei großen Datenmengen automatisch viele unterschiedliche Ausprägungen, was zu

einem schlechten Laufzeitverhalten führt. Deshalb müssen numerische Variablen in Kategorien eingeteilt werden, was einen erhöhten Aufwand bedeutet. Mittels einer geeigneten grafischen Darstellung lassen sich Entscheidungsbäume leicht interpretieren. Die gefundenen Erkenntnisse können auch als logische Ausdrücke modelliert werden und lassen sich auf neue Datensätze anwenden. Bei vielen Problemen mit einer überschaubaren Komplexität lassen sich Entscheidungsbäume schnell und einfach generieren. Allerdings nehmen bei einem Anstieg der Dimensionalität die Bäume beträchtliche Ausmaße an, was die Interpretierbarkeit erschwert. Insbesondere in den tieferen Verzweigungen wird der Einfluss von zufälligen Elementen (fehlende Werte, Ausreißer in den Daten, ...) größer, was zu einer Übermodellierung des Entscheidungsbaumes führt. Zur Umgehung des Problems bietet sich an, nur eine bestimmte maximale Tiefe der Bäume zuzulassen oder eine Mindestanzahl der Objekte pro Knoten zu fordern.

4.1.4.2 Neuronale Netze

Neuronale Netze sind aus dem Wunsch heraus entstanden, das menschliche Gehirn mitsamt seiner Lernfähigkeit nachzubilden. In Analogie zu einem Neuron im Gehirn ist das Grundelement eines Neuronalen Netzes eine Verarbeitungselement, das mehrere gewichtete Eingänge, eine Aktivierungsfunktion und einen Ausgang besitzt. Die schematische Darstellung eines Neurons ist in Abbildung 4.5 zu sehen. Die Lernfähigkeit besteht in der Anpassung der einzelnen Kantengewichte zwischen den einzelnen Neuronen. Dabei wird die Informationsverarbeitung in zwei Schritten durchgeführt:

1. Die Inputgrößen ($X_1, X_2, \ldots\ldots X_n$) werden mit ($W_1, W_2, \ldots W_n$) gewichtet und aufsummiert.
2. Der im 1. Schritt ermittelte Input führt mittels einer Aktivierungsfunktion zu einer Entscheidung, die an den Output angelegt wird.

Bei den Aktivierungsfunktionen können im wesentlichen zwei Typen unterschieden werden:

1) Bei der Klasse der Sigmoid-Funktionen wird das Neuron dann zu einer Reaktion veranlasst, wenn der Input einen gewissen Schwellwert überschreitet. Man sagt, das Neuron „feuert“. wird der Schwellwert nicht überschritten, so feuert das Neuron nicht

2) Bei der Verwendung von Radialen Basis-Funktionen feuert das Neuron nur, wenn sich der Input in der Nähe eines bestimmten Schwerpunktes befindet.

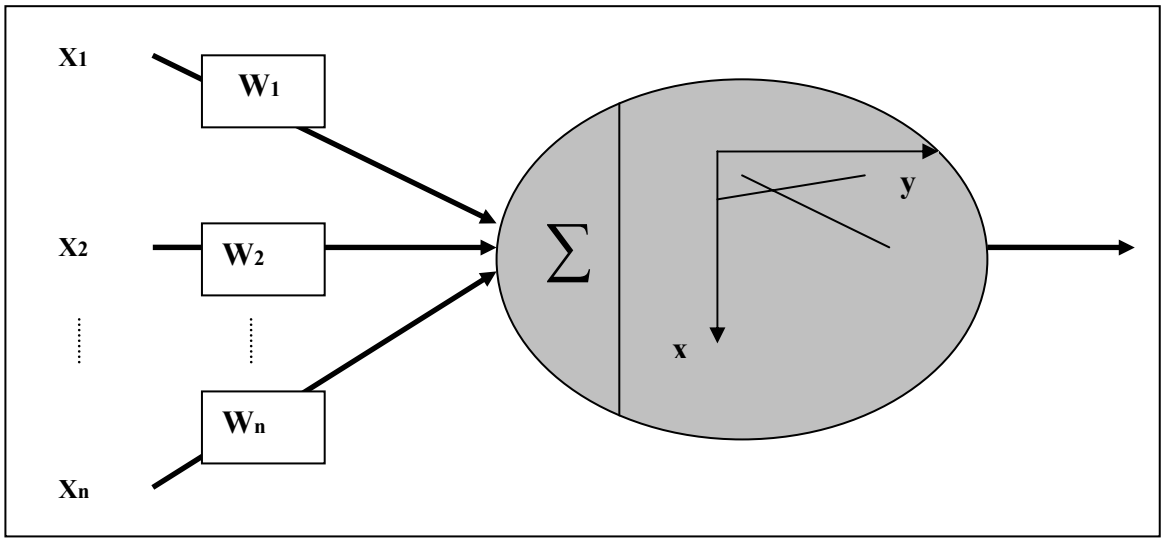


Abb.4.5: Schematische Darstellung eines Neurons

Zur Bildung eines Neuronalen Netzes werden mehrere Verarbeitungseinheiten Miteinander verbunden, üblicherweise in Ebenen, sogenannten *Layers*. Zwei Ebenen haben dabei eine Verbindung nach außen: Die Eingangsebene(*Input Layer*) und die Ausgangsebene(*Outputput Layer*). Die anderen Ebenen werden als versteckte Ebenen (*Hidden Layers*) bezeichnet. Zur Zeit werden neuronale Netze in drei Topologien eingeteilt:

1. Vorwärtsvermittelnde Netzwerke (*feed-forward*)
2. Partiell rekursive Netzwerke (*limited recurrent*)
3. Vollständig rekursive Netzwerke (*fully recurrent*)

Bei einem vorwärtsvermittelnden Netzwerk entspricht die Topologie einem azyklischen Graphen, das Netzwerk enthält also keine Zyklen. Demgegenüber sind rekursive Netze mit ihren Zyklen in der Lage, auch Zustände zu speichern.
Beim Betrieb eines neuronalen Netzes kann grundsätzlich zwischen überwachten (*supervised learning*) Lernen und unüberwachten Lernen (*unsupervised learning*) unterschieden werden:

- Überwachtes Lernen wird meist dazu verwendet, um Anwendungen zur Klassifizierung und Zeitreihenanalyse zu realisieren. Dazu werden in einer Lernphase mit Hilfe von Trainingsdaten die Verknüpfungen zwischen den Verarbeitungseinheiten

angepasst, so dass bei einem vorgegebenem Eingangssignal ein bekanntes Ergebnis eintritt. Das Lernen des neuronalen Netzes wird durch Auswertung des Fehlers im Output und durch Anpassung der Netzparameter(Input-Gewicht und Schwellwerte bzw. Schwerpunkte der Aktivierungsfunktionen) derart bewerkstelligt, dass der Fehler beim nächsten Schritt kleiner wird. Diese Lernregel durch Fehlerrückrechnung wird *Backpropagation*-Lernregel genannt. In einer zweiten Phase kann das trainierte neuronale Netz auf neue Daten angewandt werden.

- Unüberwachtes Lernen wird dann eingesetzt, wenn man eine Datenmenge untersuchen Soll und die Frage, jedoch noch nicht die Antwort formulieren kann. Dies ist vor allem bei der Datensegmentierung der Fall, wo beispielsweise Fragestellungen der Form „ Welche Kunden zeigen ein ähnliches Kaufverhalten? “ auftreten. Ein neuronales Netz versucht nun, Gemeinsamkeiten der präsentierten Eingangsmuster durch einen Ähnlichkeitsvergleich zu entdecken und seine Gewichtsstruktur danach auszurichten. Anhand dieser gefundenen Muster werden die Kunden in Zielgruppen segmentiert.

Neuronale Netze können sowohl nominal- als auch intervallskalierte Daten verarbeiten und erreichen auch bei großen Datenmengen akzeptable Laufzeiten. Es kann auch mit unvollständigen Daten gearbeitet werden, was gerade für die Mustererkennung wichtig ist. Ihr größter Vorteil liegt jedoch in der Möglichkeit unerkannte, auch nichtlineare Zusammenhänge abzubilden. Die hohe Flexibilität bezahlt man allerdings mit sehr aufwendigen und komplizierten Trainingsvorgängen. Ein Nachteil von neuronalen Netzen liegt in der fehlenden Erklärungskomponente. Es werden zwar Ergebnisse erzielt, jedoch kann der Weg, welcher zu dieser Lösung geführt hat, nicht erklärt werden. Zur Zeit bestehen auch noch keine allgemeinen Richtlinien für den Entwurf eines neuronalen Netzes, so dass sich der Modellierungsprozess (Anzahl der Schichten, Vergabe der Anfangsgewichte) auf Erfahrungswerte des Entwicklers stützt.

4.1.4.3 Assoziationsregeln

Die algorithmische Umsetzung der Analyseverfahren zur Aufdeckung von Assoziationen basiert auf der Häufigkeitsbetrachtung von Attributkombinationen. Dazu werden zwei grundlegende Maße verwendet:

- Der Träger einer Attributmenge gibt an, wie prozentual häufig die Attribute gemeinsam innerhalb des gesamten Datenbestandes vorkommen.

$$\text{Träger}(\text{Produkt A , Produkt B}) = \frac{\text{Anzahl der Transaktionen, die A und B enthalten}}{\text{Anzahl aller Transaktionen}}$$

- Die Konfidenz einer Assoziationsregel A->B gibt prozentual an, wie oft bei Zutreffen von A auch tatsächlich B zutrifft.

$$\text{Konfidenz}(\text{Produkt A -> Produkt B}) = \frac{\text{Anzahl der Transaktionen, die A und B enthalten}}{\text{Anzahl aller Transaktionen, die A enthalten}}$$

Der Prozess zur Aufdeckung von Assoziationen lässt sich dabei in zwei Phasen gliedern [Agr93]:

- In einer ersten Phase wird die Datenbank nach Teilmengen, den sogenannten *Frequent Itemsets* durchsucht, deren Häufigkeit eine minimale untere Schranke überschreitet.
- In einer zweiten Phase werden aus diesen Teilmengen die eigentlichen Assoziationsregeln generiert. Auch dabei müssen für die einzelnen Regeln bestimmte Bedingungen erfüllt sein, beispielsweise dass die Konfidenz einer Assoziationsregel größer oder gleich einer bestimmten unteren Grenze ist.

Mit diesen Vorgaben von minimalen Trägern und Konfidenzen kann gesteuert werden, ab wann eine Assoziation als interessant angesehen wird. Denn sonst treten in einer Umfangreichen Datenbasis fast beliebig viele Assoziationen auf. Dies hat auch einen positiven Einfluss auf die Rechenzeit des Verfahrens.

Im Gegensatz zu einem Entscheidungsbaum werden bei der Aufdeckung von Assoziationen hierarchische Struktur abgeleitet. Die Vorgabe eines Zielkriteriums ist nicht notwendig, vielmehr wird die Suche auf statistisch auffällige Muster beschränkt. Eine typische Anwendung dieser Methode ist die Warenkorbanalyse, bei der nach Produkten gesucht wird, die häufig gemeinsam gekauft werden. Daraus lassen sich Empfehlungen für die Regalanordnung oder Werbestrategien ableiten. Für Assoziationsregeln lassen sich aber auch leicht Anwendungen in anderen Bereichen finden,

beispielsweise im Bankgewerbe. Dabei könnten die Finanzdienstleistungen identifiziert werden, die von einem Kunden gemeinsam in Anspruch genommen wurden.
Der Vorteil von Assoziationsregeln liegt darin, dass sie leicht implementiert werden können. Sie präsentieren häufig unerwartete Resultate und tragen so dazu bei, dass das Anwendungsgebiet genauer untersucht wird, und neue Fragestellungen auftauchen. Der Nachteil liegt in einem hohen Speicherplatzbedarf bei der Erzeugung der Trägermengen und der damit verbundenen langen Laufzeiten. Außerdem können die Ergebnisse sehr umfangreich und damit unüberschaubar werden.

4.1.4.4 Clustering

Eines der am längsten eingesetzten Verfahren zur Segmentierung ist das *K-means Clustering* und wird im wesentlichen in zwei Schritten durchgeführt. Zunächst wird aus dem Datenbestand eine bestimmte Anzahl *k* von Objekten ausgewählt, die in der Folge als Clusterrepräsentanten anzusehen sind. Hierbei ist es wichtig, dass diese Auswahl möglichst repräsentativ ausfällt. In einem zweiten Schritt wird nun jedes weitere Objekt demjenigen Cluster zugeordnet, zu dessen Repräsentanten die größte Ähnlichkeit besteht. In einem iterativen Prozess werden dann die Repräsentanten angepasst und bereits zugeordnete Objekte wiederum analysiert. Insgesamt entsteht ein wechselnder Prozess der Klassifizierung und der Klassenneudefinition. Mittlerweile existieren viele Abwandlungen und Variationen des Verfahrens, wodurch eine hohe Anpassungsfähigkeit an das zu lösende Problem gegeben ist. Problematisch ist, dass die Anzahl der zu bestimmenden Cluster fest vorgegeben ist. Um eine optimale Anzahl zu bestimmen müssen mehrere Durchläufe mit verschiedenen Werten für *k* durchgeführt und die Ergebnisse miteinander verglichen werden. Eine Erweiterung des Verfahrens ergibt sich durch die Anwendung von Fuzzy-Logik. Hierbei werden keine scharfen, sondern prozentuale Segmentszugehörigkeiten betrachtet. Das Verfahren heißt *Fuzzy-k-means*. Ein weiteres, seit langem bekanntes Verfahren ist das hierarchische Clustering. Hierbei wird eine ganze Hierarchie möglicher Segmentierungen ermittelt. Es gibt zwei Vorgehensweisen:

- Jedes Objekt repräsentiert zunächst eine eigene Klasse. Es werden nun sukzessive Klassen miteinander verschmolzen, bis nur noch ein Cluster übrig bleibt. Dieses Vorgehen wird „agglomeratives hierarchisches Clustering“ genannt.
- Jedes Objekt gehört zunächst einem Cluster an. Es werden danach neue Teilcluster erzeugt, bis nur noch Cluster bestehend aus Einzelobjekten übrigbleiben.

Bei diesem Vorgehen, welches „divises hierarchisches Clustering“ heißt, werden also bestehende Cluster in neue Cluster aufgeteilt.
Im Gegensatz zum k-means Clustering kann eine optimale Lösung aus dem Gesamtbild heraus gefunden werden, da durch die verschiedenen Hierarchiestufen mehr als eine Lösung präsentiert wird. Nachteilig ist, dass wegen der fehlenden Iteration anfangs getätigte Falschklassifizierungen nicht mehr rückgängig gemacht werden können.
Bei den Verfahren des Data Mining unterscheidet man prinzipiell zwischen überwachtem und unüberwachtem Lernen(vgl. im Abschnitt 4.1.4.2). Im überwachten Fall ist die Klassenzugehörigkeit der Objekte bereits bekannt. Diese Information kann sowohl für die Generierung als auch die Evaluierung von Klassifikatoren verwendet werden. Im unüberwachten Fall dagegen sind eine derartige Klasseneinteilung sowie die Anzahl der Klassen a priori völlig unbekannt. Durch die Anwendung von Clusteranalyse soll vielmehr eine eventuell vorhandene Klassenstruktur erst gefunden werden. Nachfolgend kann diese Klassenstruktur dann überwachter Verfahren(z.B. Entscheidungsbaumverfahren) als Eingabe zur Verfügung gestellt werden.

4.1.4.5 Statistische Verfahren

Wie schon bei den Entscheidungsbäumen und Clusteranalyse gesehen, können die statistischen Verfahren nicht vollständig vom Data Mining getrennt werden. Vielmehr sind sie ein elementarer Bestandteil, welcher zum einen klassische Analyseverfahren für Data Mining bereitstellt und zum anderen die Voraussetzungen für die Anwendbarkeit der modernen Verfahren schafft. Die Regressionsanalyse bildet dabei eines der flexibelsten und am häufigsten eingesetzten statistischen Analyseverfahren. Sie dient zur Analyse von Beziehungen zwischen einer abhängigen Variablen (Regressand) und einer oder mehreren unabhängigen Variablen (Regressoren). Insbesondere wird sie eingesetzt, um Zusammenhänge zu erkennen und zu erklären und die Werte der abhängigen Variablen zu schätzen bzw. zu prognostizieren. Dabei wird eine lineare Beziehung zwischen den Variablen unterstellt, d.h. dass aus den empirischen Werten für Regressand und Regressor(en) eine lineare Beziehung errechnet wird, die durch den folgenden allgemeinen Ausdruck dargestellt werden kann:

$$y = \beta_0+\beta_1 x_1+\beta_1 x_2 \ldots . +\beta_n x_n$$

y = Regressand

β_0 = konstantes Glied

β_n =Regressionskoeffizient des n-ten Regressors

x_n =n-ter Regressor

Generell liegt der Vorteil der Regressionsansätze darin, dass sie auf einer fundierten Theorie Basieren, die dem Analytiker einen differenzierten Einblick in die Mechanismen des Modells ermöglicht. Hinzu kommt, dass die für die Modellierung verfügbaren numerischen Algorithmen sehr weit entwickelt sind und deshalb sehr schnell arbeiten.

Das Problem der linearen Verfahren ist, das sie komplexe Zusammenhänge oft zu einfach interpretieren, d.h. bestimmte Grundannahmen entsprechen nicht unbedingt der Realität. Zur Modellverbesserung können nichtlineare Verfahren genommen werden, jedoch erschweren die höhere Anzahl der Parameter und komplexere Funktionsformeln die Handhabung. Es sind mehr manuelle Eingriffe notwendig als bei den linearen Methoden, wodurch die Laufzeiten erhöht werden. Auch Visualisierungstechniken spielen bei der Mustererkennung und der Entdeckung von Abweichungen eine große Rolle. Im wissenschaftlich-analytischen Bereich ist das Streudiagramm der am häufigsten verwendete grafische Darstellungstyp. In diesem werden die Merkmalsträger durch Punkte gekennzeichnet, deren Lage in Bezug zu einem rechtwinkligen Koordinatensystem beschrieben wird. Dieses Koordinatensystem wird aus der Ordinatenachse (Y-Achse) und der Abszissenachse (X-Achse) gebildet. Besitzt ein Merkmalsträger für das Zeilenmerkmal X die Merkmalsausprägung x und für das Spalten Merkmal Y die Merkmalsausprägung y, so wird der ihm zugeordnete Punkt durch das Koordinatenpaar (x, y) gekennzeichnet. Allgemein verwendbar und in einer Vielzahl von Varianten verfügbar, kann mit einem Streudiagramm Art, Richtung und Intensität des Zusammenhangs zwischen zwei Merkmalen dargestellt werden. Darüber hinaus ist eine Identifikation ungewöhnlicher Beobachtungswerte sowie das Erkennen von Gruppenbildungen unter den Beobachtungswerten leicht möglich.

Streudiagramme können um zusätzliche Informationen ergänzt werden, die das Aufdecken von Mustern erleichtern. Vor allem bei einer großen Anzahl von Beobachtungswerten oder variierender Streuung ist es empfehlenswert, das

Streudiagramm mit weiteren Informationen anzureichern, die Hinweise auf die zugrunde liegende Struktur geben können.

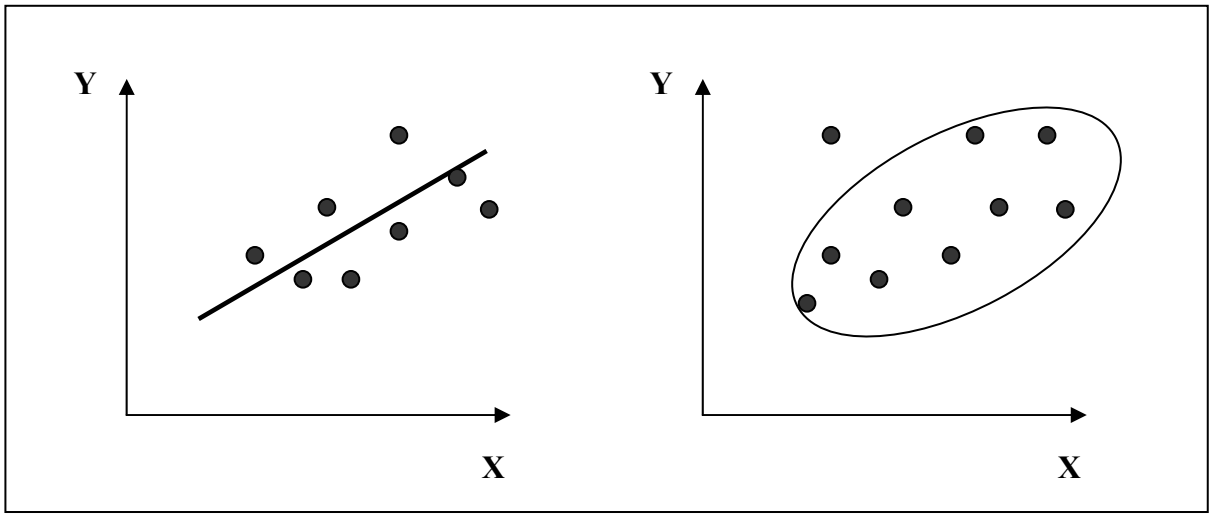


Abb.4.6: Streudiagramm-Varianten mit Regressionsgeraden und Dichte-Ellipse

Das einfachste Beispiel für derartige Verfahren ist das Einzeichnen einer Regressionsgeraden in die Punktwolke (siehe Abbildung 4.6), welches genau die Gerade ist, die von der Gesamtheit der Punkte den geringsten Abstand hat. Eine weitere nützliche Anwendung ist Dichte-Ellipsen (für 95% bzw. 99% des Datenmaterials) in das Streudiagramm.

4.2 OLAP

In diesem Abschnitt werden zunächst die Begriffe Online Transactional Processing (OLTP) und Online Analytical Processing (OLAP) voneinander abgegrenzt. Anschließend werden einige moderner Analysewerkzeuge kurz dargestellt, die zur Identifikation von Wissensstrukturen führen können.

4.2.1 Abgrenzung zu OLTP

Im operativen Tagesgeschäft werden heute in nahezu allen Wirtschaftszweigen, zur Verwaltung der umfangreichen und ständig wachsenden Datenbeständen Datenbanksysteme eingesetzt. Sie sorgen für eine dauerhafte und fehlerrobuste Speicherung sowie einen effizienten Zugriff. Dieser Art des Einsatzes lässt sich mit dem Begriff

des Online Transactional Processing (OLTP) umschreiben [Bre98]. Typische Beispiele hierfür sind Anwendungen wie „Buchung eines Fluges“ in einem Flugreservierungssystem oder „Verarbeitung einer Bestellung“ in einem Handelsunternehmen. OLTP-Anwendungen zeichnen sich dadurch aus, dass sie nur relativ kleine Datenmengen pro Transaktion zu verarbeiten haben und dass sie nur auf dem jüngsten, aktuell gültigen Zustand des Datenbestandes operieren. Die typische Datenstruktur besteht meist aus einem relationalen Datenbanksystem mit flachen, nicht hierarchischen Tabellen.

4.2.2 Grundlagen

Während sich OLTP-Anwendungen auf die Verarbeitung von Transaktionen konzentrieren, benötigen viele Managementaufgaben Informationen aus unternehmensinternen Datenquellen, jedoch in einer aufbereiteten und verdichteten Form. Die dabei verwendeten Anwendungen greifen auf „historische“ Daten zurück, um daraus z.B. Rückschlüsse auf die Entwicklung des Unternehmens zu ziehen. Typische Anfragen bei einer Fluggesellschaft oder einem Handelsunternehmen sind:

- Wie hat sich die Auslastung der Transatlantikflüge über die letzten zwei Jahre entwickelt?
- Wie haben sich besonders offensive Marketingstrategien für bestimmte Produktlinien auf die Verkaufszahlen ausgewirkt?

Hier interessieren nicht mehr einzelne Geschäftsvorfälle, wie sie in den operationalen Systemen des Tagesgeschäftes abgebildet sind, sondern managementrelevante und Aufbereitete Kenngrößen. Für diese Anforderungen sind die OLTP-Systeme mit ihren Strukturen nicht konzipiert worden. Darüber hinaus besteht in relationalen Systemen bei umfangreichen Datenbeständen oft eine unüberschaubare Anzahl von Tabellen, Views und Joins, deren Struktur nur von einem versierten Anwender beherrscht werden kann.

Aus diesem Grund stellte E.F. Codd im Jahr 1993 anhand von 12 Regeln einen neuen Ansatz vor, das Online Analytical Processing. Er definiert OLAP als … the name given to the dynamic enterprise analysis required to create, manipulate, animate und synthesize information from ‚Enterprise Data Models’. This includes the ability to discern new or unanticipated relationships between variables, the ability to identify the parameters necessary to handle large amounts of data, to create an unlimited

number of dimensions (consolidation paths) and to specify cross-dimensional conditions and expressions.[Cha99]. Unter der angeführten Definition ist ein Forderungskatalog zu verstehen, der als Maßstab für all OLAP-Produkte gelten sollte. Im Mittelpunkt dieses Ansatzes steht hierbei vor allem die multidimensionale Sicht auf die Daten. Damit können die gespeicherten Datenbestände auf vielfältige Art und Weise und von verschiedenen Seiten betrachtet werden. Aufgrund spezialisierter Abfragesprachen, einfacher Navigation im multidimensionalen Datenbestand, intuitiver Benutzungsoberflächen und geeigneter Ergebnispräsentation bleibt die tatsächliche Komplexität einer Abfrage dem Anwender verborgen. Zur Veranschaulichung der Multidimensionalität der Datenbasis wird diese oft als dreidimensionaler Datenwürfel abgebildet, da in einer Graphik nicht mehr Dimensionen darstellbar sind. Tatsächlich unterliegt sie aber keiner derartigen Beschränkung. Durch eine intuitive Datenbearbeitung, z.B. die Rotation des Würfels (*Slicing*) oder das Auswählen einer Teilmenge (*Data Dicing*), besteht für den Anwender eine erhöhte Flexibilität , so dass neue Ideen und Hypothesen schnell überprüft werden können. Sollen detaillierter Fragestellungen beantwortet werden, so kann der Verdichtungsgrad der einzelnen Dimension verändert werden. Will man beispielsweise den Jahresumsatz auf die einzelnen Monate abbilden so spricht man von einem *Drill down*, da dadurch eine weniger starke Verdichtung der Daten stattfindet. Bei einem Wechsel zu einer höheren Verdichtungsstufe spricht man von einem *Roll up.*

Man kann drei Arten von OLAP-Anwendungen unterscheiden [[Cha99],[Bre98]):

- MOLAP – Multidimensionales OLAP:

Das Multidimensionale OLAP ist die klassische OLAP- Architektur. Die im RDBMS oder anderen Quellen vorliegenden Daten werden in die OLAP-Datenbank in Form einer optimierten, mehrdimensionalen Struktur geladen, ganz oder teilweise kalkuliert und stehen dann zur Abfrage mit einem Frontend bereit.

Für MDBMS existiert im Gegensatz zu relationalen Datenbanken kein Standard. Dies gilt sowohl für Datenformate, Abfragesprache und den Aufbau der Metadaten. Ein Datenaustausch zwischen den verschiedenen Systemen oder die Verwendung herstellerfremder Tools oder Front-Ends ist dadurch nicht ohne weiteres möglich Die multidimensionale Datenbank bietet einen schnelleren Zugriff als eine relationale Datenbank und sind aus technischer Sicht nicht so skalierbar wie eine relationale Datenbank.

- ROLAP – Relationales OLAP:

Die zur Analyse notwendigen Daten verbleiben im RDBMS und brauchen nicht zusätzlich geladen werden. Die zur Abfrage notwendigen SQL-Anweisungen werden zur Laufzeit erzeugt und auf die Transaktions- oder Aggregationstabellen abgesetzt. Relationales OLAP kann dadurch eine größere Anzahl von Dimensionen und Ausprägungen verarbeiten, kann dem Benutzer aber keine verlässliche Aussage über die zu erwartenden Antwortzeiten liefern.

- HOLAP – Hybrides OLAP:

Um die Vorteile beider Technologien ROLAP und MOLAP zu verbinden, werden oftmals hybride Systeme realisiert. Dabei werden die Vorteile der relationalen Technik (z.B. hohe Skalierbarkeit) mit der unerreichten schnellen Zugriffszeit bei multidimensionalen Datenbanken verknüpft. In der Regel wird dabei aus einer bestimmten Sicht des multidimensionalen Würfels auf die Detaildaten (im relationalen System) zugegriffen. Der Vorteil liegt darin, dass bei diesem Zugriff Indices optimal ausgenutzt werden können. Der Nachteil liegt in der aufwendigeren Realisierung.(die Komplexität des Gesamtsystems zu erhöhen)

Die wichtigsten Unterschiede zwischen MOLAP und ROLAP werden in der Tabelle 1 zusammengefasst.

OLAP-Methode	**MOLAP**	**ROLAP**
Vorteile	Platzsparende Speicherung mehrdimensionaler Daten.	Zuverlässige, langjährig erprobte Technologie, mit der auch sehr umfangreichen Datenmengen verwaltet werden können.
	Schneller Zugriff durch auf mehrdimensionale Daten optimierte OLAP-Server.	Zugriff über den SQL-Standard. Auch die Erstellung der Meta-Daten ist in einem gewissen Umfang standardisiert (DDL).
	Berechnung und Speicherung der Aggregationen erfolgt multidimensional.	Reichhaltiges Angebot an Management-Werkzeugen und Spezialisten für die Verwaltung eines RDBMS.
	MOLAP-Server sind relativ einfach in Betrieb zu nehmen und erfordern nur einen geringen Wartungsaufwand (bedingt allerdings durch wenige Einstellmöglichkeiten)	Teilweise flexibler, wenn sich der Aufbau von Dimensionen regelmäßig ändert. Auch können Dimensionen in einigen Produkten automatisch mit den operationalen Daten synchronisiert werden.
Nachteile	Keine anerkannten Standards für	Relationale Strukturen sind

	Abfragesprache, Datenformate, Meta-Daten oder Programmierschnittstellen.	ungeeignet zur Speicherung mehrdimensionaler Daten. Dies führt bei ROLAP zu höherem Platzbedarf und tendenziell geringerer Performance.
	Geringere Flexibilität, wenn sich der Aufbau der Dimensionen ändert. Oft müssen dann große Teile oder sogar die komplette OLAP-Datenbank neu berechnet werden.	Die Inbetriebnahme eines RDBMS ist zeitaufwendig, der Wartungsaufwand oft enorm.
	Generell bieten MOLAP-Server nur wenige Einstellmöglichkeiten für Performance-Optimierungen, Lastverteilung, etc.	

Tabelle 1: Gegenüberstellung der Vor- und Nachteile von MOLAP und ROLAP (in Anlehnung an[Pop99]).

4.3 Information Mapping

Dieser Abschnitt stellt eine Dokumentationsart als die gängige Methode vor. Information Mapping ist eine Darstellungsmethode für Dokumentationen und Richtlinien. Im folgenden wird die Technik vorgestellt.

4.3.1 Einführung

Der Einsatz der Informationstechnologie in vielen Bereiche dient dazu, die Bereitstellung der Informationen zu erhöhen, die Zugriffzeit auf die Informationen zu verkürzen und das Austausch der individuellen, kollektiven Informationen in verteilten Organisationen zu ermöglichen, führt aber auch dazu, dass die Informationsangeboten in Organisationen rapid zunimmt und das Auffinden von Informationen zunehmend schwieriger wird. Ein Grund liegt darin, dass die Informationen in Dokumenten nicht wissensorientiert strukturiert sind. So ist die zielgerichtete Informationsstrukturierung in Dokumenten eine wesentliche Voraussetzung für die Identifikation von Informationen und deren Wiederverwendbarkeit [vgl. Hol00]. Eine Methode, die sich genau dies zum ziel gesetzt hat, ist Information Mapping. Information Mapping ist eine weltweit erfolgreich angewandte Methode für das strukturierte Verfassen von Information. Texte und Dokumente werden eindeutig

und übersichtlich - wie in einer "Map" - einer Karte. In folgende werde ich einen Überblick über diese Methode geben.

4.3.2 Was ist Information Mapping

Die Information Mapping Methode wurde von Prof. Robert E. Horn (Harvard University, Boston) zur Vereinfachung des Erstellungsprozesses von Dokumenten und zur Unterstützung der Informationsaufnahme entwickelt. Sie basiert auf den Erkenntnissen der Kognitionspsychologie im Bereich der menschlichen Wahrnehmung und Informationsverarbeitung und hilft, Information so aufzuschreiben und zu strukturieren, dass sie von einem Leser optimal verarbeitet werden können. So wird Information Mapping wie folgendes definiert :

“Information Mapping ist eine Methode, gedruckte Informationen so aufzubereiten, dass die Anordnung von Text (und Grafiken) auf einer Seite bereits die Struktur und die Beziehungen der einzelnen Informationen abbildet. Die unterschiedlichen Anordnungen entsprechen dabei den Beziehungen und Abhängigkeiten der Informationen [Imm99]“.

Nach dieser Definition wird die Forderung dieser Methode verdeutlicht, mit welcher Informationen analysiert, gegliedert, dargestellt und identifiziert werden können. Diese Methode stellt uns modulare Einheiten und Bausteine sowie Prinzipien zur Verfügung:

- Es gibt vordefinierte Informationseinheiten
- Informationsarten werden als Informationskategorien behandelt
- Es gelten sieben Prinzipien

Diese Punkte erlauben uns, einen Text so zu verfassen, dass sowohl die Bedürfnisse des Lesers als auch die des Verfassers berücksichtigt werden.

4.3.3 Die Grundelementen der Information-Mapping-Methode

4.3.3.1 Informationseinheiten

Die Information Mapping Methode führt zwei klar definierte Informationseinheiten ein.

- den Block

Ein Block besteht aus einem oder mehreren Absätzen, Tabellen und Diagramm. Er enthält nur eine in sich abgeschlossene Art von Information und ist stets betitelt.

- Die Map

Eine Gruppe von mehreren zusammenhängenden Blöcken(siehe Abb.4.7), die unter Beachtung der Prinzipien der Methode zusammengefügt werden, wird als Map Bezeichnet. Ein Dokument besteht aus mehreren Maps (Ein Beispiel befindet sich im Anhang).

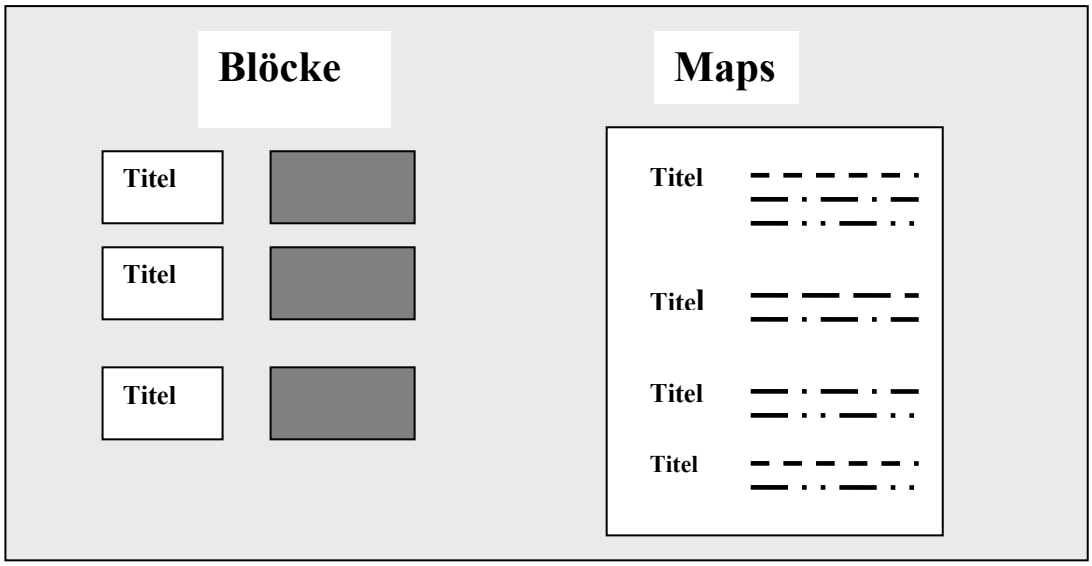


Abb.4.7 : Die Informationseinheiten nach der Information-Mapping-Methode[Hol00]

4.3.3.2 Informationstype

Informationstype sind grundlegende Kategorien, welche aufgrund ihres Zweckes oder ihrer Aussage für den Leser eingeordnet werden können. Die Information Mapping Methode unterscheidet zwischen Prozedur, Prozess, Struktur, Begriff, Prinzip, Fakten und Klassifikation. Für jede dieser Informationsarten gibt die Methode mögliche Vorgehen an, damit der Autor einfacher einen Block gestalten kann.

- Eine *Prozedur* ist eine Folge von Schritten, welche von einer Person ausgeführt werden,
- um ein bestimmtes Resultat zu erzielen.
- Ein *Prozess* ist eine zeitliche Abfolge von Ereignissen oder Phasen, mit dem Ziel ein bestimmtes Resultat zu erhalten.

- *Strukturen* beschreiben den Aufbau von Gegenständen oder Objekten, die klare Umrisse haben und gegliedert werden können.
- Unter *Begriff* versteht man eine Klasse oder Gruppe von Dingen mit gemeinsamen Merkmalen.
- Ein *Prinzip* ist ein Grundsatz, nach welchem der Leser sein Handeln auszurichten hat.
- *Fakten* sind Tatsachen, Daten oder Ereignisse.
- Unter *Klassifikation* versteht man die Gliederung von Begriffen oder Bezeichnungen nach bestimmten Merkmalen.

Die nachfolgende Tabelle zeigt den Strukturensunterschiede in Dokumente zwischen der gängigen und der Information-Mapping-Methode.

Gängige Methode	**Information-Mapping**
Titel	Titel
Kapitel	Maps
Unterkapitel	Block
Abschnitt	Prozedur, Prozess, Struktur, Begriff, Prinzip Fakten oder Klassifikation
Satz, Grafik oder Tabelle	Satz, Grafik oder Tabelle

Tabelle 2. Strukturensunterschiede zwischen der gängigen und IMAP-Methode

4.3.3.3 Prinzipien

Information Mapping definiert auch sieben Prinzipien, die der Strukturierung der Informationen helfen.[vgl. Hol00]:

1. *Gliederung:*

Das Gliederungs-Prinzip gliedert die gesamte Informationsmenge zu einem begrenzten Thema in übersichtliche und leicht zu verarbeitende Informationseinheiten.

2. *Relevanz-Prinzip:*

Gemäß Relevanz-Prinzip enthält jede Informationseinheit nur Informationen, die zusammengehören und die in Bezug auf ihren Zweck oder Aussage für den Leser relevant sind.

3. *Betitelung:*

Das Betitelungs-Prinzip besagt, dass jede Informationseinheit mit einem Titel versehen ist, der klar den Zweck, die Aussage, die Funktion oder den Inhalt der Information bezeichnet.

4. *Einheitlichkeit:*

Gemäß dem Einheitlichkeits-Prinzip werden vergleichbare Informationsarten oder Sachgebiete einheitlich behandelt.

5. *Gleichwertigkeit:*

Das Gleichwertigkeits-Prinzip besagt, dass Grafiken, Zeichnungen, Bilder und Tabellen direkt in den Text eingefügt werden sollen.

6. *Verfügbarkeit*

Gemäß dem Verfügbarkeits-Prinzip werden dem Leser alle Informationen dort zur Verfügung gestellt, wo er sie auch braucht.

7. *systematische Gliederung*

Das systematische Gliederungs-Prinzip besagt, dass kleine, in sich geschlossene Informationseinheiten zu größeren Einheiten zusammengefasst werden.

Das Erstellen der Dokumente orientiert sich in erster Linie an den Bedürfnissen des Lesers und beantwortet zwei wesentliche Fragen, die sich jener im Hinblick auf die von ihm verlangte Tätigkeit stellt:

- Wie muss ich vorgehen?
- Was muss ich dazu wissen?

Aus der oben angeführten Darstellung von Information Mapping ergibt es sich, dass der wichtigste Aspekt eines Themas die Kern- oder Schlüsselinformation ist. Sie erfahren, wie Sie diesen „Kern“ identifizieren können, wo in einer Map er untergebracht werden sollte und wie Sie ihn wirkungsvoll präsentieren können. Die „Schlüsselinformation“ wird im sogenannten Schlüsselblock aufgearbeitet. Sie ist der Kern einer Map. Ein praktisches Beispiel ist im Anhang A (in Anlehnung an[Test]) zu sehen. Dort wird die Darstellung eines Dokuments im Nomaltext mit der Darstellung im Information Mapping Text verglichen. Daraus erkennt man sich sofort, dass die Informationen, die in dem mit Information-Mapping-Methode erstellten Dokument erneut erfasst sind. leicht überschaubar und auffindbar sind, Das liegt an die besseren Strukturierung, den kleinen Informationseinheiten und an der Übersichtlichkeit des gemappten Texts.

4.4 RDF

Die Entwicklung der einfach zu bedienender Web-Browser führt zu einem rapiden Anwachsen der Daten am Web und damit auch der zur Verfügung stehenden Informationen. Die Menge der über das Internet zugreifbaren Informationen wächst weiter an und es wird zunehmend schwieriger, relevante Information aufzufinden. Gegenwärtige Katalogsysteme und Suchmechanismen genügen den Anforderungen der Benutzer nach nachgefragter Information und zuverlässigem Wissen nicht.

Es gilt neue Strategien aus dem gegenwärtigen Standard und zukunftsorientierte Technologien zu entwickeln. Es gilt viele Fragen zu klären: "Wo bekomme ich Informationen zu einem bestimmten Thema? Habe ich die richtige (relevante) Information? Wie ist die Qualität der erhaltenen Information? Wie kann ich eine bestimmte Information wieder finden?" Die Softwareentwickler müssen auf diese Fragen Antworten finden , um Millionen von Benutzern Werkzeuge zu geben, damit diese die Information finden, nach der sie suchen. Die Lösung wird in der Angabe von Metadaten über bestimmte Informationsressourcen gesehen [Las99].

4.4.1 Matadaten

Unter Metadaten ("Daten über Daten") versteht man strukturierte Daten, mit deren Hilfe eine Informationsressource beschrieben und dadurch besser auffindbar gemacht wird. Der Begriff geht zwar dem Web-Zeitalter voraus, findet aber vor allem im Zusammenhang mit modernen elektronischen Informationssystemen seine Anwendung. Von World Wide Web Consortiums (W3C) stammt die Definition: "Metadaten sind maschinenlesbare Informationen über elektronische Ressourcen oder andere Dinge."

Mit sehr großen Datenmengen sinnvoll umgehen zu können, ist insbesondere beim Internet eine wichtige aber schwierige Aufgabe. Entscheidende Bedeutung dafür besitzen die Metadaten: die Daten über Daten. Nur wenn die Metadaten in ausreichender Tiefe erfasst und maschinenlesbar bzw. -verständlich abgespeichert und verarbeitet werden können, entsteht ein ausreichender Nutzen für den Anwender.

Metadaten liefern also Grundinformationen über ein Dokument, wie z.B. Angaben über Autor, Titel oder Zeitpunkt der Veröffentlichung, und reproduzieren damit im

Prinzip genau das, was an Erschließungsarbeit in den Bibliotheken seit jeher geleistet wurde. Hinter dem Begriff der Metadaten steht deshalb auch die Suche nach neuen Ansätzen in der Ressourcenbeschreibung und nach den entsprechenden Verfahren der Informationsvermittlung, die auf einen effizienten und kostengünstigen Einsatz in elektronischen Netzen hin optimiert sind. Insgesamt erhebt die gegenwärtige Metadatendiskussion den Anspruch, dass damit bessere Erschließungs- und Retrievalmechanismen angeboten werden können, als sie bisher im Internet existieren. Und da sich ein immer größerer Bereich des Informationsangebots z. B von Bibliotheken aus diesem Medium wie auch aus anderen elektronischen Quellen speist, wird die gesamte Diskussion um Metadaten von zunehmender Bedeutung.
Die effektive Nutzung von Metadaten erfordert eine allgemeinen Konvention für die Semantik, die Syntax und die Strukturen, um Informationen aus den Daten zielgericht aufzufinden und daraus das neue Wissen zu generieren.
Das Resource Description Framework (RDF) - eine XML-Applikation entwickelt vom World Wide Web Consortium (W3C) 1999 - legt die Grundlagen, um solche Metadaten verarbeiten zu können. Nachfolgend werden das RDF-Modell , die RDF-Syntax und RDF-Schema kurz skizziert.

4.4.2 RDF-Datenmodell

Das RDF-Modell basiert auf wohletablierten Prinzipien der Datenrepräsentation. Wie aus dem Namen „Resource Description Framework" hervorgeht, spielt der Begriff Ressource dabei eine zentrale Rolle. Die Grundlage von RDF ist ein Modell zur Repräsentation benannter Eigenschaften und Werte dieser Eigenschaften. Damit übernimmt RDF die bekannte Philosophie von Attribut-Value-Paaren.
Das RDF-Datenmodell besteht aus folgenden 3 Objekt-Typen:

- Ressourcen:

Jedes beliebige Ding kann eine Ressource darstellen. Dabei können Ressourcen z.B. eine Webpage, Teile einer Webpage, eine Sammlung von Webpages, aber auch Objekte sein, auf die gar nicht über das Web zugegriffen werden kann, z.B. Bücher, CDs, Geräte und allgemein jede vorstellbare Einheit, die durch eine URI (Uniform Ressource Identifier) identifiziert werden können. Ressourcen in RDF sind also mit Topics(siehe im Abschnitt 4.5) vergleichbar.

- Eigenschaften:

Eine Eigenschaft (engl. "Property") ist, wie der Name schon sagt, eine spezifische Charakteristik oder ein spezifisches Attribut, das der Beschreibung einer Ressource dient. Jede Eigenschaft definiert erlaubte Werte und Beziehungen mit anderen Eigenschaften.

- Aussagen:

Eine Aussage (engl. "Statement") ist eine spezielle Ressource zusammen mit einer Eigenschaft dieser Ressource und dem Wert dieser Eigenschaft. Diese 3 Teile einer Aussage nennt man auch Subjekt, Prädikat und Objekt. Das Objekt (also der Wert einer Eigenschaft) kann entweder eine weitere Ressource oder ein Literal (siehe Abb.4.8)

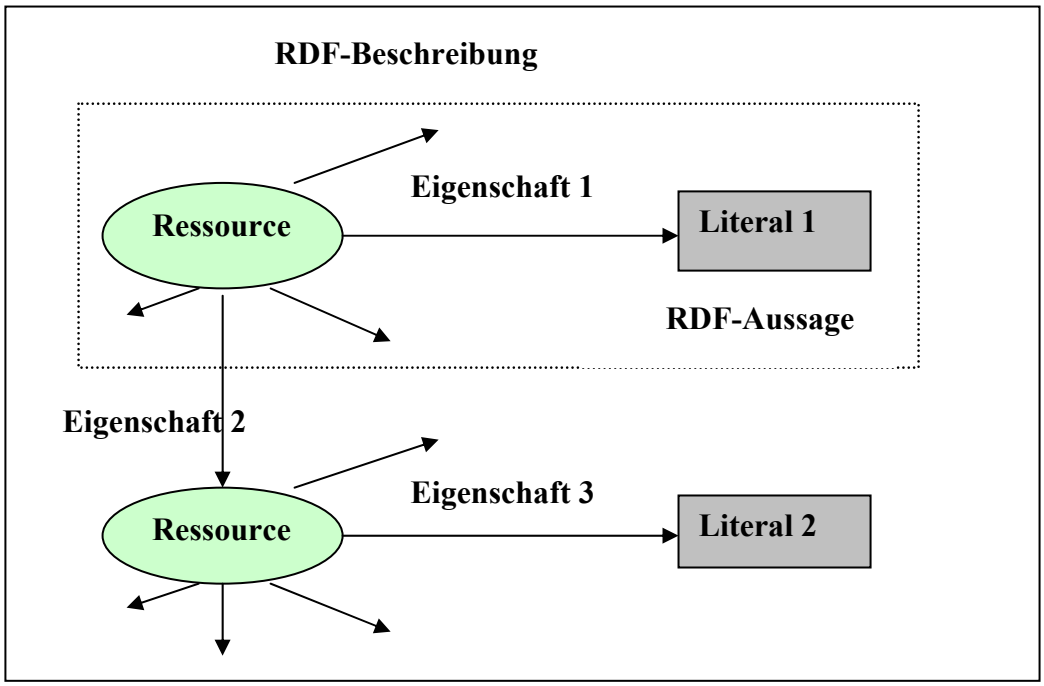


Abb.4.8: Zusammenhänge zwischen den einzelnen Objekttypen

Einer Ressource können Eigenschaften (properties) zugeordnet werden können (z.B. Name oder Adresse einer Person oder einer Web-Seite) und Eine Eigenschaft kann durch ein Literal ausgedrückt kann und somit als Attribut der Ressource betrachtet werden und Literale können atomare Werte wie Unicode-Zeichenketten oder Zahlen sein, können aber auch beliebige strukturierte Daten beinhalten. Neben Literalen kann eine Eigenschaft auch eine Beziehung mit einer weiteren Ressource herstellen. Eine Menge von Eigenschaften einer Ressource wird als RDF-Beschreibung bezeichnet

Ein Beispiel aus dem WWW wäre z.B. der Autor einer Webpage:

Der Autor der Seite http://www.uni-paderborn.de/~shengxi/index.html ist Shengxi Long

Hier wäre:

Das Subjekt(die Ressource)	*http://www.uni-paderborn.de/~shengxi/index.html*
Das Prädikat(die Eigenschaft)	Autor
Das Objekt(Literal)	„Shengxi Long“

Zur Veranschaulichung solcher RDF-Modelle werden gerichtete Graphten verwendet:

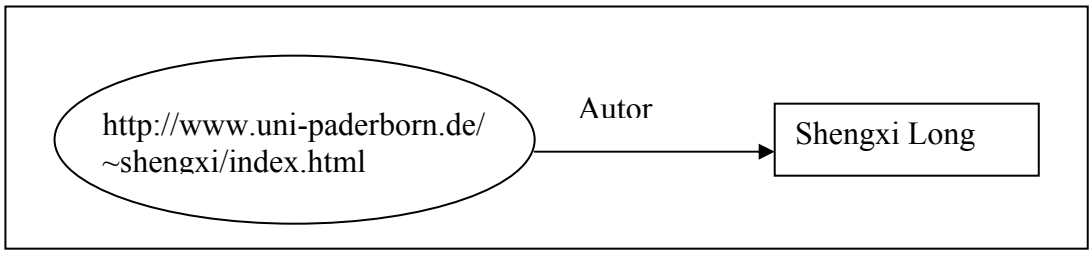


Abb.4.9: Beispiel für ein RDF-Datenmodell

dabei ist die Richtung der Kante wichtig. Sie geht immer vom Subjekt zum Objekt, sodass das oben dargestellte Diagramm die obige Aussage entspricht.
Das RDF-Datenmodell gestattet die Beschreibung von anderen Beschreibungen. Somit ist es z.B. möglich auszudrücken, dass das Institut für Wirtschaftsinformatik 2 verkündigt, dass die Vorlesung Office-System von Prof. Nastansky gehalten wird". Damit solche Aussagen modelliert werden können, müssen Aussagen als Ressourcen behandelt werden. Die ursprüngliche Aussage als Ressource können mit vier vordefinierten Eigenschaften modelliert wird:

- *rdf:subject* bezeichnet die Ressource, die das Subjekt der Aussage darstellt (Vorlesung Office-System im obigen Beispiel).
- *rdf:predicate* identifiziert das Prädikat der Aussage (gehalten-von).
- *rdf:object* bezeichnet den Wert der Eigenschaft in der ursprünglichen Aussage (im obigen Beispiel die Ressource, die dem Individuum Prof. Nastansky entspricht).
- *rdf:type* beschreibt den Typ der Ressource, die die Aussage repräsentiert. Alle durch solchen Aussagen entstandenen Ressourcen sind vom Typ rdf:Statement.

Eine detaillierte Beschreibung der Ausdrucksfähigkeit des RDF-Datenmodells sowie die vollständige RDF-Grammatik sind in [RDF00] zu finden.

4.4.3 RDF-Syntax

Zur Aufbereitung und zum Austausch von Daten wird eine konkrete Syntax benötigt. Durch die Verwendung von XML können Instanzen des RDF-Modells in maschinen- und menschenlesbarer Form abgespeichert und ausgetauscht werden. Dieser Prozess wird als Serialisieren bezeichnet. Die RDF-Spezifikation bietet für die Serialisierung zwei XML-Syntax Typen an.
Während die Standard (Serialization) Syntax eine Kodierung in einer sehr „gesetzmäßigen“ Weise zur Verfügung stellt, zielt die Kompakte(Abbreviated) Syntax auf eine kürzere und kompaktere Schreibweise ab. Die folgende Abbildung zeigt die Verwendung der RDF-Syntax unter Benutzung des RDF-Namespaces und des DC (Dublin core) -Namespaces:

```
<?xml version="1.0"?>
<rdf :RDF
        xmlns:rdf = „ http://www.w3.org/1999/02/22-rdf-syntax-ns# “
        xmlns:s=“ http://description.org/schema/ “
        <rdf:Description about ="http://www.uni-paderborn.de/~shengxi/index.html">
                <DC:Creator>shengxi Long</DC:Creator>
</rdf :Description>
</rdf :RDF>
```

Abb.4.10: Serialization Syntax von RDF in XML

In der ersten Zeile steht der so genannte Prolog mit dem natürlich jedes XML-Dokument begonnnen wird. Danach folgt das eigentliche RDF-Element, welches die Beschreibung (Descripion) enthält. Das XML-Element *rdf:RDF* signalisiert, dass dieses Element eine RDF-Beschreibung enthält. *rdf:Description* fasst eine Reihe von Aussagen über die Ressource zusammen, die durch das Attribut *about* spezifiziert wird. Die Angabe der beiden Namespaces bedeutet, dass alle Eigenschaften der

Beschreibung aus einer der beiden Namespaces stammen und der RDF-Namespace den Default-Wert darstellt.

Neben der vorgestellten Serialization Syntax wurde auch noch eine zweite Form Entwickelt. Dies hat folgende Gründe:

- um die Lesbarkeit zu steigern und
- um Problemen mit älteren Browsern, welche Teile der RDF-Beschreibung (insbesondere jene mit Start- und Endtags) im Browserfenster darstellen würden, vorzubeugen.

Bei dieser Notation werden die Eigenschaften einer Ressource auf Attribute ihres umgebenden Elements abgebildet. Somit ergibt sich für das in Abbildung 4.11 dargestellte Modell folgende Notation:

```
<?xml version="1.0"?>
<rdf :RDF
        xmlns:rdf = „ http://www.w3.org/1999/02/22-rdf-syntax-ns# “
        xmlns:s=“ http://description.org/schema/ “
        <rdf :Description about ="http://www.uni-paderborn.de/~shengxi/
            index.html">
            DC:Creator=”shengxi Long “/>
</rdf :RDF>
```

Abb.4.11: Abbreviated Syntax

Um eine Ressource, die mehrere Eigenschaften vom gleichen Typ besitzt(z.B.: kann eine Ressource mehrere Autoren haben), zu beschreiben, bietet RDF Container-Konstrukte an:

Dabei unterscheidet man 3 verschiedene Arten von Containern:

Bag: eine ungeordnete Liste von Ressourcen oder Literalen, wobei die Reihenfolge der Elemente keine Rolle spielt, z. B. Liste Autoren eines Buches.

Sequence: eine geordnete Liste von Ressourcen oder Literalen, wobei die Reihenfolge der Elemente wichtig ist, z.B. chronologische Liste aller Präsidenten der USA.

Alternative: eine Liste möglicher alternativen Elementen, von denen nur eine ausgewählt werden kann z. B. Liste von Mirror-Seiten im Internet.

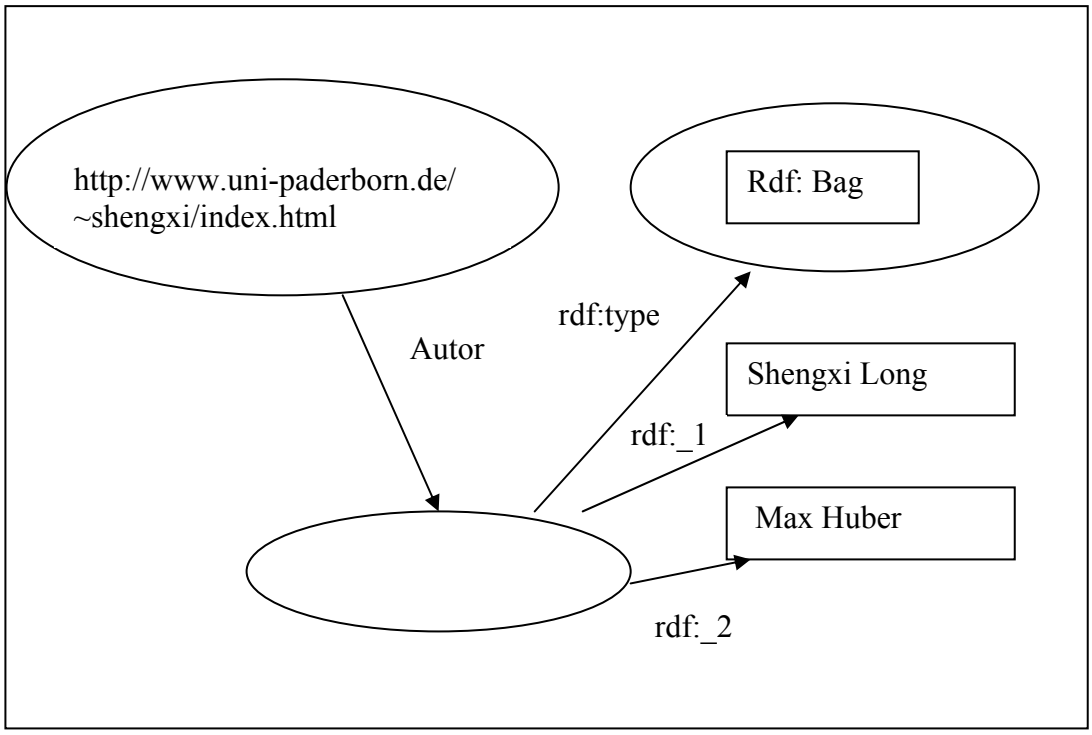


Abb.4.12: Bag-Container

Die Element solcher Liste werden mit Namen(„_1“, „_2“, usw.) innerhalb der Container durchnummeriert (siehe Abb.4.12). Eine solche Notation ist bei Verwendung der Abbreviated (kompakten) Syntax erforderlich. Alternativ kann bei der Serialization (Standard) Syntax vor jedes Listenelement das „rdf:li“-Tag gesetzt werden, sodass die Nummerierung automatisch von oben nach unten vorgenommen wird. Somit ergibt sich durch die Serialisierung des in Abb.4.13 dargestellten Graphen folgender Code:

```
<?xml version="1.0"?>
<rdf:RDF
     xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
     xmlns:s="http//description.org/schema/"
    <rdf:Description about=2http://www.uni-paderborn.de/~shengxi/index.html">
        <DC:Creator>
                <rdf:Bag>
                        <rdf:li>shengxi Long</rdf:li>
                       <rdf:li>Max Huber</rdf:li>
                </rdf:Bag>
        </DC:Creator>
     </rdf:Description>
</rdf:RDF>
```

Durch RDF wird es aber auch möglich, die Daten aus Datenbanksystemen in ein universelles Format zu exportieren. Auch hier besteht der wesentliche Vorteil darin,

dass bei RDF-Daten durch das dazugehörige Schema eine semantische Interpretation der Daten ermöglicht wird.

4.4.4 RDF-Schema

Das RDF-Datenmodell erlaubt eine präzise, intersubjektiv und maschinell Nachvollziehbare logische Erfassung der Informationen. Es bietet jedoch keine Hilfe bei der Wahl der Grundbegriffe und -beziehungen, die bei der Beschreibung bestimmter Sachverhalte benötigt werden. Somit können Anwendungen zwar genau feststellen, worüber gesprochen wird, wissen aber nicht, was damit gemeint wird. Dieses Verständnis soll durch RDF-Schemas erreicht werden. RDF-Schemas beschreiben Begriffsontologien, die in der Regel anwendungsspezifisch sind. Damit wird ein Typisierungssystem für RDF-Datenmodell durch die Möglichkeit zur Definition von Klassen und Unterklassen. Ressourcen zur Verfügung gestellt können, dann Instanzen einer solchen Klasse sein. RDF-Schemata sind im Grunde genommen Entity-Relationship.

Für das korrekte Verständnis der RDF-Aussagen ist die eindeutige Interpretation der Bedeutung seiner Bestandteile, d.h. der Prädikate und der beteiligten Ressourcen, ausschlag-gebend. In einem globalen Medium wie dem WWW kann man sich nicht darauf verlassen, da z.B. alle potentiellen Kommunikationspartner mit dem Begriff „Dozent" eine identische Bedeutung verknüpfen. Damit die Kommunikation nicht scheitert, sollte man so präzise wie nur möglich sein. RDF-Schemas sind nichts anderes, als ein „Wörterbuch", in dem eine Menge von Begriffen samt ihrer Gebrauchsanweisung beschrieben werden kann. Der Namensraum-Mechanismus von XML erlaubt es, jedem Begriff das zugehörige Schema zuzuordnen. Eine menschen- und maschinenlesbare Beschreibung des entsprechenden RDF-Schemas kann durch Auflösung der Namensraums-URIs gewonnen werden

In dem Dokument [RDFS99] wird nicht ein Schema definiert, in dem verschiedene Klassen und Eigenschaften festgelegt werden, sondern es wird eine *"Schema Definition Language"* definiert, mit deren Hilfe die eigentlichen Schemas definiert werden. Diese eigentlichen Schemas werden auch als *Vokabulare* bezeichnet, da sie definieren, welche Eigenschaften mit welcher Bedeutung für Beschreibungen zu Verfügung stehen.

Im folgenden sollen einige wesentliche Punkte der "Schema Definition Language" erläutert werden, die vollständige Beschreibung ist in [RDFS99] zu finden.

4.4.4.1 Klassen und Eigenschaften

eine Ressource kann eine Instanz einer oder mehrerer Klassen sein, was durch die Eigenschaft *rdf:type* ausgedrückt wird. Die Bildung von Klassenhierarchien ist ebenfalls möglich, dazu kann die Eigenschaft *rdfs:subClassOf* verwendet werden.

In [RDFS99] werden außerdem verschiedene Ressourcen definiert, die es ermöglichen, Aussagen über Beschränkungen für die Verwendung von Eigenschaften und Klassen zu machen. So wird zum Beispiel ermöglicht, gültige Werte für Eigenschaften zu definieren und festzulegen, welchen Klassen eine Eigenschaft sinnvollerweise zugeordnet werden kann. Allerdings wird kein Mechanismus festgelegt, der diese Beschränkungen durchsetzt. Ob und wie die Beschränkungen durchgesetzt werden, bleibt den einzelnen Anwendungen überlassen.

Folgende grundlegende Klasse werden definiert:
rdfs:Resource ist die Klasse aller Ressourcen, d.h. alle Dinge, die mit RDF beschrieben werden, sind Instanzen dieser Klasse

rdfs:Property ist die Klasse der Eigenschaften. Sie repräsentiert eine Untermenge der Ressourcen.

rdfs:Class korrespondiert zum generischen Konzept eines Typs, alle Klassen sind Instanzen des Typs Klasse (auch die Klasse *rdfs:Class* selbst). Ein ähnliches Konzept existiert z.B. in Java. Beachtenswert ist hier, dass einerseits *rdfs:Class* eine Unterklasse von *rdfs:Resource* ist, da die Klassendefinitionen selbst Ressourcen sind, andererseits *rdfs:Resource* auch eine Instanz von *rdfs:Class* ist, da es eine Klasse ist.

Außerdem werden die folgenden grundlegenden Eigenschaften festgelegt:
rdf:type ermöglicht die Zuordnung einer Ressource zu einer Klasse, die Ressource ist dann Instanz der Klasse und es kann dann angenommen werden, dass sie der typischen Charakteristik der Klasse entspricht.

rdfs:subClassOf spezifiziert eine Untermengen/Obermengen-Relation zwischen Klassen. Die Relation ist transitiv. Ressourcen, die Instanzen einer bestimmten Klasse sind, sind auch Instanzen von deren Oberklassen.

rdfs:subPropertyOf ist eine Instanz von *rdf:Property* und wird benutzt um anzugeben, daß eine Eigenschaft eine Spezialisierung einer oder mehrerer anderer Eigenschaften ist. Wenn eine Eigenschaft E2 Spezialisierung einer anderen Eigenschaft E1 ist, und

eine Ressource hat eine Eigenschaft E2 mit dem Wert B, dann wird damit implizit ausgedrückt, dass sie auch eine Eigenschaft E1 mit Wert B hat.

Beispiel: Wenn biologicalFather eine Spezialisierung von biological Parent ist, und Max ist der biologicalFather von John, dann ist Max auch ein biological Parent von John.

Mit *rdf:seeAlso* kann auf eine andere Ressource verwiesen werden, die Informationen über die vorliegende Ressource enthält.

rdfs:isDefinedBy ist eine Spezialisierung von *rdfs:seeAlso* und verweist auf eine Ressource, die die vorliegende Ressource definiert.

4.4.4.2 Beschränkungen

In einem RDF-Schema können den Klassen und Eigenschaften Beschränkungen (*Constaints*) zugeordnet werden. Insbesondere werden damit die Konzepte von Domain und Range umgesetzt. Die Domain einer Eigenschaft gibt die Klassen an, auf die die Eigenschaft sinnvollerweise angewendet werden kann. Die Range legt die gültigen Werte für eine Eigenschaft fest. So könnte z.B. festgelegt werden, daß die Werte einer Eigenschaft "Autor" nur vom Typ "Person" sein dürfen und die Eigenschaft nur bei Ressourcen vom Typ "Buch" angewendet werden darf.
Ein RDF-Datenmodell, welches eine Beschränkung verletzt, wird als *inkonsistent* bezeichnet. Das RDF-Schema benutzt die Beschränkungs-Eigenschaften *rdfs:range* und *rdfs:domain*, um festzulegen, in welcher Weise seine Eigenschaften benutzt werden dürfen. Beispielhaft sollen hier drei Elemente erläutert werden:

rdfs:comment wird verwendet, um eine natürlich sprachliche Beschreibung einer Ressource anzugeben. Es kann also als Eigenschaft jeder Ressource verwendet werden, der Wertebereich sind alle gültigen Literale.

rdfs:label dient zur Angabe einer lesbaren Beschriftung für eine Ressource. Die Charakteristik ist analog zu rdfs:comment.

rdfs:range gibt, wie bereits erwähnt, den gültigen Wertebereich einer Eigenschaft an. Es ist somit als Eigenschaft jeder Eigenschaft verwendbar und der Wertebereich sind sämtliche Klassen.
Im Abschnitt 1 wird zunächst die Sprache Englisch für das Dokument angegeben und es werden Namespaces für die RDF-Syntax und das RDF-Schema festgelegt.

Abschnitt 2 definiert eine Klasse Person, die Unterklasse der Klasse Animal ist. Als natürlich sprachliche Beschreibung wird "The class of people." angegeben.

Der Abschnitt 3 definiert eine Eigenschaft age, die auf Instanzen der Klasse Person zutrifft, der gültige Wertebereich der Eigenschaft sind Werte vom Typ Integer.

Im Abschnitt 4 wird eine Eigenschaft maritalStatus definiert, die ebenso auf Instanzen der Klasse Person angewendet werden kann. Der gültige Wertebereich sind dabei Instanzen der Klasse MaritalStatus, die in Abschnitt 5 definiert wird. Instanzen der Klasse werden dann in Abschnitt 6 definiert.

```
(1) <rdf:RDF xml:lang="en"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/TR/1999/PR-rdf-schema-19990303#">

(2) <rdfs:Class rdf:ID="Person">
     <rdfs:comment>The class of people.</rdfs:comment>
     <rdfs:subClassOf
      rdf:resource="http://www.classtypes.org/useful_classes#Animal"/>
   </rdfs:Class>

(3) <rdf:Property ID="age">
     <rdfs:range
      rdf:resource="http://www.datatypes.org/useful_types#Integer"/>
     <rdfs:domain rdf:resource="#Person"/>
   </rdf:Property>

(4) <rdf:Property ID="maritalStatus">
     <rdfs:range rdf:resource="#MaritalStatus"/>
     <rdfs:domain rdf:resource="#Person"/>
   </rdf:Property>

(5) <rdfs:Class rdf:ID="MaritalStatus"/>

(6) <MaritalStatus rdf:ID="Married"/>
   <MaritalStatus rdf:ID="Divorced"/>
   <MaritalStatus rdf:ID="Single"/>
```

```
<MaritalStatus rdf:ID="Widowed"/>
</rdf:RDF>
```

4.5 Topic Maps

Der neue ISO Standard Topic Maps eröffnet ungeahnte Erschließungsmöglichkeiten der immer unüberschaubarer werdenden Informationsflut. Betitelt als "GPS des Informationsuniversums" vereinfacht er die Strukturierung und damit die zielgerichtete Suche und Navigation in großen Datenmengen. Im folgenden Kapitel werden Teile des grundlegenden Standards ISO/IEC 13250 beschrieben.

4.5.1 Einführung

Mit wachsenden organisationalen Wissensbasen wird der Einsatz neuer Informationstechnologien immer notwendiger. Die Problematik, mit der sich Topic Maps beschäftigen, ist nun die Modellierung bzw. Realisierung einer intelligenten Navigation für Informationsressourcen in organisationalen Intranets wie auch im Web. Dabei wird besonderes Augenmerk auf die Beziehungen zwischen den Begriffen bzw. Themen und auf die Quellenverweise der Begriffe bzw. Themen gelegt. Denn durch den Aufbau solcher Beziehungen lassen sich logische Schlussfolgerungen ziehen, wodurch neues Wissen entstehen kann.

4.5.2 Ein Beispiel zur Problemdarstellung

Angenommen eine Fragestellung für eine Suche im Web lautet: „Welche Opern wurden von deutschen Komponisten komponiert, die von Mozart beeinflusst waren? “ Für diese Fragestellung würde keine Suchmaschine eine zufrieden stellende Antwort finden, d.h. konkrete Links auf Webseiten, die dieses Thema behandeln, ausgeben. Man würde allein schon an der Eingabe der Fragestellung scheitern, welche womöglich folgendermaßen formuliert sein könnte: „Oper + Deutschland + Komponist + Mozart“. Die Erfahrung hat jedoch gezeigt, dass dies nur wenig Sinn und noch weniger befriedigende Resultate ergibt. Diese Tatsache beruht darauf, dass Suchmaschinen, auch unter der Hinzunehme von Metadaten, nur Webseiten liefern, in deren Inhalt diese Schlagwörter zwar vorkommen, jedoch nicht in der Lage sind

festzustellen, ob eine logische Beziehung zwischen ihnen im Kontext besteht. Schließlich wäre es doch viel praktischer, direkt nach Opern suchen zu können, die von Komponisten komponiert worden sind und diese Komponisten mit Deutschland und Mozart in Verbindung setzen zu können („geboren in“, „beeinflusst von“). Darüber hinaus sollten neben den Ergebnissen der Suche (in diesem Fall sind das die Namen der Opern) auch Referenzen auf Informationsquellen zu den Ergebnissen gegeben sein[Rth].

4.5.3 Grundlegende Begriffserklärung

Topic Maps behandeln also eine intelligente Navigation in Wissensstrukturen. Solch ein semantische strukturiertes Netzwerk besteht aus einer Zusammentragung von Informationsobjekten und deren Verknüpfung untereinander. Jeder Knoten in diesem Netzwerk stellt die zielgerichtete Informationsstrukturierung in Dokumenten einen Topic dar. Im obern angeführten Beispiel wären dies z.B. „Mozart“ oder „Deutschland“. Knoten werden durch logische Verweise verbunden, welche die Beziehungen (Assoziationen) zwischen den Topics repräsentieren. In Bezug auf das Beispiel sind dies „geboren in“ oder „beeinflusst von“. Ein Knoten bzw. Topic ist meist mit einer Informationsquelle (z.B. Artikel, Bild, Video, etc.) verknüpft. Verweise auf diese Ressourcen nennen sich Occurrences (Vorkommnis). Die Ressourcen befinden sich in der Regel außerhalb der Topic Map und werden über Adressierungsmethoden angesprochen. Diese ist nur eine einführende Erklärung. Genauer werden die Elemente von Topic Maps im betreffenden ISO-Standard erläutert.

4.5.4 ISO-Standard Topic Maps

Im Herbst 1999 ist der ISO-Standard Topic Maps (ISO/IEC 13250: 1999) verabschiedet worden. Dieser Standard beschreibt die einzelnen Komponenten einer Topic Map. Die Elemente repräsentieren die Topics, die Assoziationen zwischen Topics und die Verweise auf die Informationsquellen. Als konzeptionelle Bausteine stehen Topics (mit Topic Types und Topic Names), Topic Occurrences so wie Topic Associations bereit. Weiterführende Konzepte sind Scopes, Public Subject Descriptor und Facets. Alle diese Konzepte werden im folgende einzeln beschrieben.

4.5.4.1 Topics

Ein Topic kann vieles darstellen, z.B. eine Person, einen Gegenstand, ein Land, ein Wort, eine Zahl, etc. Es ist eine Art Container, der erst mit Inhalt gefüllt werden muss. Dieser Inhalt charakterisiert dann das Topic und ist anwendungsspezifisch. Je nach Anwendungsgebiet kann ein Topic dann einen Begriff (z.B. in einem Lexikon), eine Komponente (z.B. in der technischen Dokumentation), ein Projekt (z.B. im Wissensmanagement) beschreiben.
In Bezug auf das obige Beispiel könnten nun folgende Topics dargestellt werden:

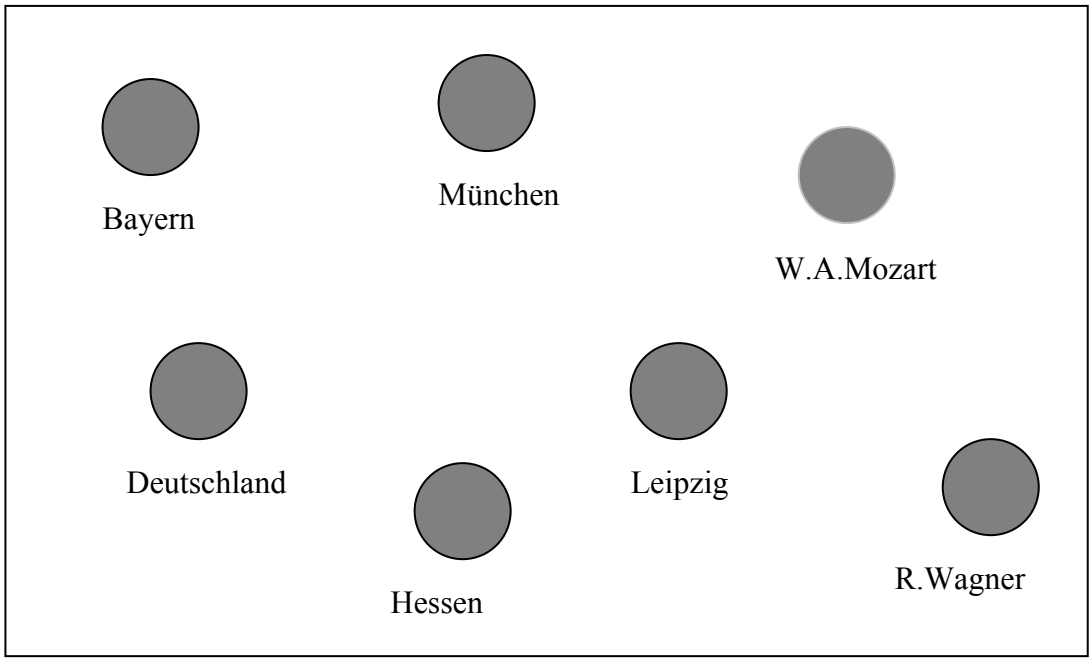


Abb.4.13: Beispiel für Topics

4.5.4.2 Topic Types

Ein Topic kann nun einen oder mehrer Typen haben. R. Wagner könnte vom Typ Person sein, Bayern und Hessen vom Typ Bundesland und Deutschland vom Typ Staat. R. Wagner könnte allerdings auch vom Typ Komponist sein und Komponist wieder vom Typ Person. Die Typen eines Topics (oder der Typ) sind selbst wieder Topics, die demnach auch deklariert werden müssen, bevor sie für andere Topics als Typ dienen können, Somit kann eine Typhierarchie gebildet werden. Vom Anwendungsfall abhängig kann es auch sein, dass ein Topic gar keinen Typ hat. Wenn etwa sonst keine Komponisten in der Topic Map vorkommen, ist es womöglich gar nicht nötig, den Topic des Komponisten einzuführen.

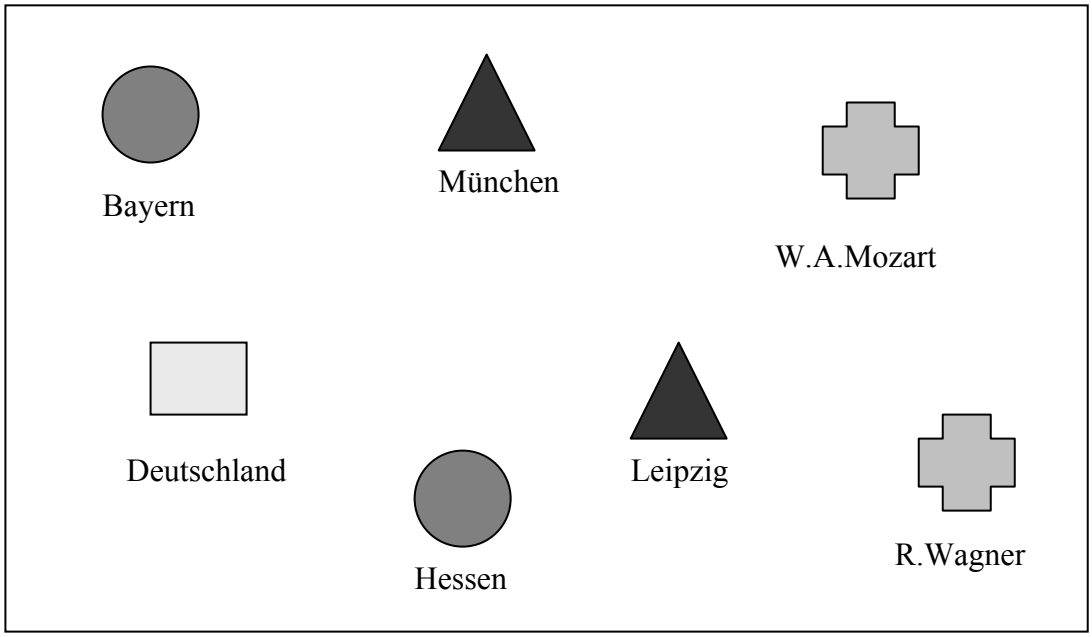


Abb. 4.14: Topics in Topic Types

4.5.4.3 Topic Names

Der ISO-Standard versucht, alle Arten von Namen für Topics zu ermöglichen und bietet dazu folgende drei Namensarten:

Base Name: Der Base Name ist der eigentliche Name eines Topics. Jeder Topic muss mindestens einen Base Name haben, um gemäß der DTD gültig zu sein.
Display Name: Der Display Name ist eine Zeichenkette, die zur Ausgabe eines Topic Names herangezogen wird. Die Angabe eines Display Names ist optional. Wird kein Display Name angegeben, übernimmt der Base Name die Rolle.
Sort Name: Der Sort Name wird zur Einordnung des Topics in sortierten Listen oder Dokumenten herangezogen. Die Angabe des Sort Names ist wiederum optional. Wird er weggelassen, übernimmt der Base Name seine Rolle.

Der Standard sieht vor, für ein und dasselbe Topic ein oder mehrere Topic Names vergeben zu können. Dies kann genutzt werden, um bei der Modellierung der Topic Map zum Beispiel die Mehrsprachigkeit zu berücksichtigen. Aber auch andere Einsatzmöglichkeiten sind denkbar. So ist beispielsweise auch die Berücksichtigung von Adelstiteln in einer Geschichtsdatenbank möglich. Ein vorstellbare Beispiel wird auch in Abb.4.15 dargestellt.

Abb. 4.15: Beispiele für Topic Names

4.5.4.4 Occurrences

Jedes Topic kann mit einer beliebigen Anzahl von Informationsressourcen verknüpft werden. Diese Ressourcen stellen Verweise bzw. Verbindungen zu externen Ressourcen dar. Jede Occurrence kann eine bestimmte Rolle übernehmen. Occurrence Roles sind ebenfalls Topics und müssen zuvor deklariert werden.

Beispielsweise könnte der Topic München eine Occurrence auf einem Stadtplan im Web beinhalten, als Occurrence Role könnte dann etwa Karte oder Stadtplan gewählt werden. Die Occurrence Role weist also der Occurrence eine gewisse Semantik zu und beschreibt somit ihre Bedeutung. In Abb.4.16 stellen die Linien die Occurrences zu den jeweiligen Web-Dokumenten dar. Die verschiedenen Arten der Linien deuten auf die unterschiedlichen Occurrence Roles hin. Aus der Abbildung wird auch die Trennung zwischen Topics und dem Informationspool(dem WEB) deutlich. Diese Trennung ist einer der Schlüssel zur Mächtigkeit des Standards. Dieser Sachverhalt wird in den nächsten Abschnitten noch ausführlicher erläutert.

Diese Ressourcen sind in ihrer Art durch den Standard nicht näher spezifiziert [Bie00]. Dadurch könnten zum Beispiel auch Dokumente, die nicht in elektronischer Form vorliegen, referenziert werden. Diese Eigenschaft lässt sich besonders gut für einen schrittweisen Umstieg im innerbetrieblichen Wissens- und Dokumentenmanagement verwerten, da auch Papierdokumente über ein geeignetes Adressierungsverfahren in der Topic Map über einen entsprechenden Verweis abgelegt werden können. Durch die Angabe eines Rollennamens wird die Kategorie festgelegt, zu der die Informationsressource gehört. Das optionale Typ-Attribut verweist auf ein Topic der Map und beschreibt dadurch den Typ der Ressource näher. Für den Adressierungsmechanismus werden durch den Standard ebenfalls keine Vorgaben gemacht.

Bei der Adressierung elektronischer Dokumente wird XLink zur Zeit verwendet. Derzeit gibt es jedoch noch nicht genügend Software, die solche Dokumente verarbeiten kann. Eine Implementation von XLink wird von Empolis (www.empolis.co.uk) mit dem Programm X2X angeboten. Die gleiche Firma arbeit auch an einer Topic Map Engine. Für die Adressierung nicht-elektronischer Informationsressourcen müsste eine anwendungsspezifische Auszeichnung dieser Ressourcen gefunden werden (zum Beispiel würde sich für eine Bibliothek die übliche Standortsignatur anbieten).

Durch die Angabe von Rollen und Typen bieten Topic Maps eine deutlich erweiterte Funktionalität gegenüber herkömmlichen Indexen in Büchern [Pep99]. Viele

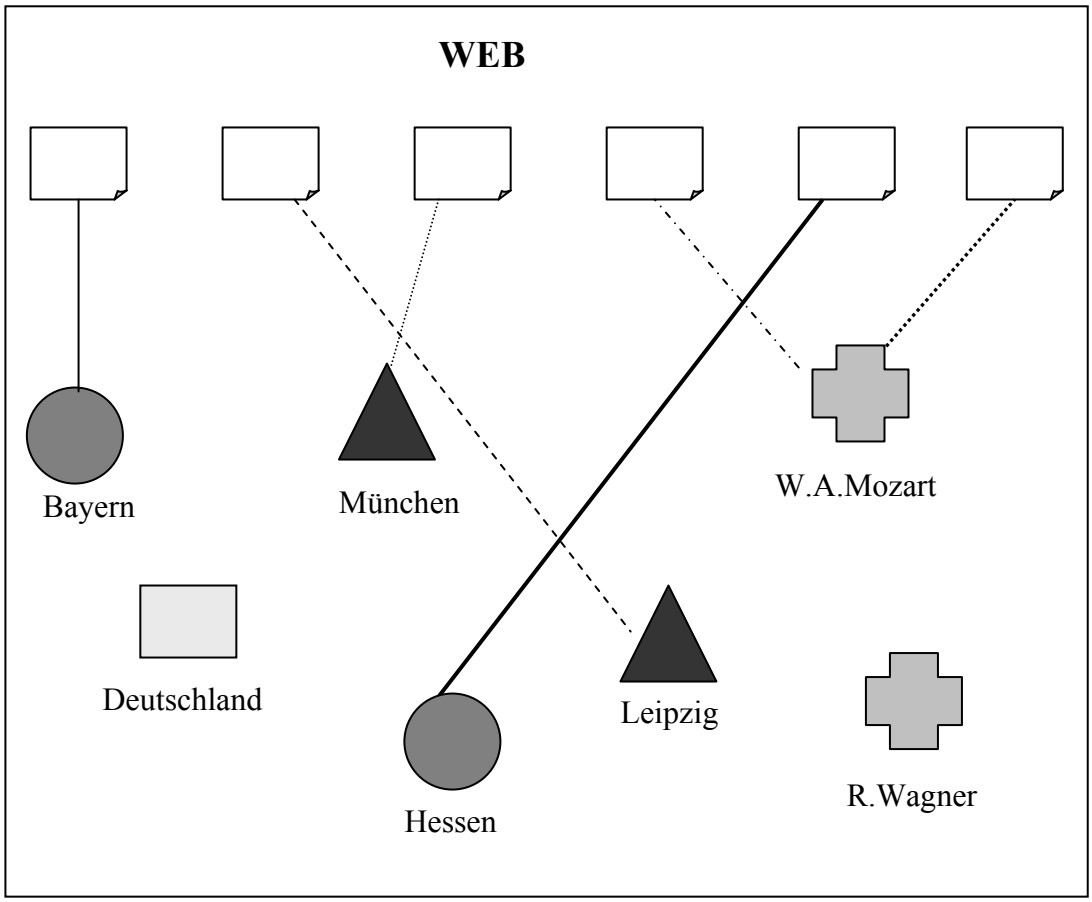


Abb.4.16: Occurrences

Bücher erweitern Indexe durch typographische Besonderheiten (z.B. Fettdruck der Seitenangabe, wenn zu einem Stichwort eine Abbildung existiert). Durch den Einsatz von Topic Maps sind solche Erweiterungen ebenfalls möglich, bieten aber dem Nutzer weitaus mehr Möglichkeiten. Durch die Angabe der *Occurrence Types* erhält der Nutzer zum Beispiel mehr Informationen zum Kontext des aufgeführten Eintrags im Index. Er kann dadurch erkennen, ob es sich bei der angegebenen Informations-

ressource um ein Video, einen Text, ein Bild oder ähnliches handelt. Der Rollenname gibt Auskunft darüber, in welchem Zusammenhang die Ressource mit dem gewählten Thema steht.

4.5.4.5 Associations

Die bisher vorgestellten Konzepte von ISO 13250 ermöglichen dem potentiellen Nutzer lediglich eine lose Sammlung von Beschreibungen von Topics sowie die Angabe von wichtigen Informationsquellen. Aber erst durch die Zuordnung von Beziehungen zwischen einzelnen Topics eröffnen sich viele neue Anwendungs-möglichkeiten.. Mit Hilfe von Assoziationen können jedoch weit mehr Informationen in eine Topic Map gepackt werden. Zum Beispiel:

- Bayern liegt in Deutschland.
- Bayern grenzt an Hessen.
- München liegt in Bayern.
- R. Wagner wurde in Leipzig geboren.
- Lohengrin wurde von Wagner komponiert.
- R. Wagner wurde von W. A. Mozart beeinflusst.

Eine Association kann jedoch maximal eine Typ (Association Type) haben, der wiederum ein Topic ist. In unserem Beispiel wären dies unter anderem *liegt in, geboren in, grenzt an oder beeinflusst von*[siehe Abb.4.17].
Associations können symmetrisch sein, etwa wie *grenzt an*. Wenn Bayern an Hessen grenzt, so muss auch Hessen an Bayern grenzen. Sie können aber auch transitiv sein. Wenn z.B. München in Bayern liegt und Bayern in Deutschland, so muss auch München in Deutschland liegen. Jedes Topic, das zu einer Association gehört, kann eine Association Role haben, die wiederum als Topic deklariert werden muss. So kann etwa bei *Bayern liegt in Deutschland* der Topic *Bayern* die Rolle *Bundesland* und *Deutschland* die Rolle *Staat* haben.
Durch die Angabe von Rollen werden die einzelnen Teile der Assoziation näher beschrieben. Die optionalen Typangaben verweisen ihrerseits ebenfalls wieder auf ein entsprechendes Topic in der Map. Die Angabe der Rollen, die einzelne Topics in einer Assoziation einnehmen, ist äußerst wichtig. Verständnisprobleme können zum Beispiel dann auftreten, wenn mehrere Topics in einer Assoziation vom gleichen Typ

sind. Es ist unzureichend zu wissen, dass ein Topic „Max Maier“ mit dem Topic „Paul Müller“ in einer „ist-Mitarbeiter-von“ Beziehung steht. Die Rolle der jeweiligen Person ist ungeklärt (wer ist Mitarbeiter?, wer Vorgesetzter?). Deshalb besitzen die Rollenangaben eine große Bedeutung für den Nutzer einer Map. Topics mit ihren zugehörigen Assoziationen sind gut vergleichbar mit semantischen Netzen. Dabei stellen die Topics die Knoten im Netz dar, während die Assoziationen die Kanten repräsentieren. Durch die Attribute kann ebenfalls der Typ der Kanten abgebildet werden [Fre00]. Topic Maps besitzen durch die bereits oben angesprochene Trennung von den Informationsressourcen einen hohen Informationswert und können als eine Art portables semantisches Netz weitergegeben werden. Ein anderer Bearbeiter dieser Map ist dann in der Lage, auch andere Informationsquellen entsprechend seinen Bedürfnissen zu nutzen (z.B. könnte eine Topic Map alle Werke eines Künstlers referenzieren, während eine zweite alle Arbeiten von Dritten [z.B. Kritikern] referenziert). Die Verbindungen zwischen den einzelnen Topics bleiben dabei jedoch erhalten.

4.5.4.6 Public Subject Descriptor

Ein Topic verfügt über ein eindeutiges Identifikationsattribut, dessen Wert als Public Subject Descriptor bezeichnet wird. Beim Zusammenfügen zweier Topic Maps werden Topics, deren Identifikationsattribute übereinstimmen, zu einem Topic zusammengefügt. Wenn also zwei Topic Map Autoren ihre Topic Maps zusammenfügen und beispielsweise beide einen Topic für Deutschland verwenden, der in einem Fall den Namen *Deutschland* und im anderen Fall den Namen *Germany* hat, doch in beiden Fällen der internationale Ländercode als Identifikationsattribut herangezogen wird, so wird daraus ein Topic mit beiden Namen entstehen. Hierbei ergibt sich jedoch ein Problem. Denn nicht für alle Topics existiert ein so eindeutiger Public Subject Descriptor wie dies bei einem Staat (internatonaler Ländercode) der Fall ist. Und selbst wenn einer existiert, so heißt dies noch lange nicht, dass er auch von jedem Topic Map Autor verwendet wird.

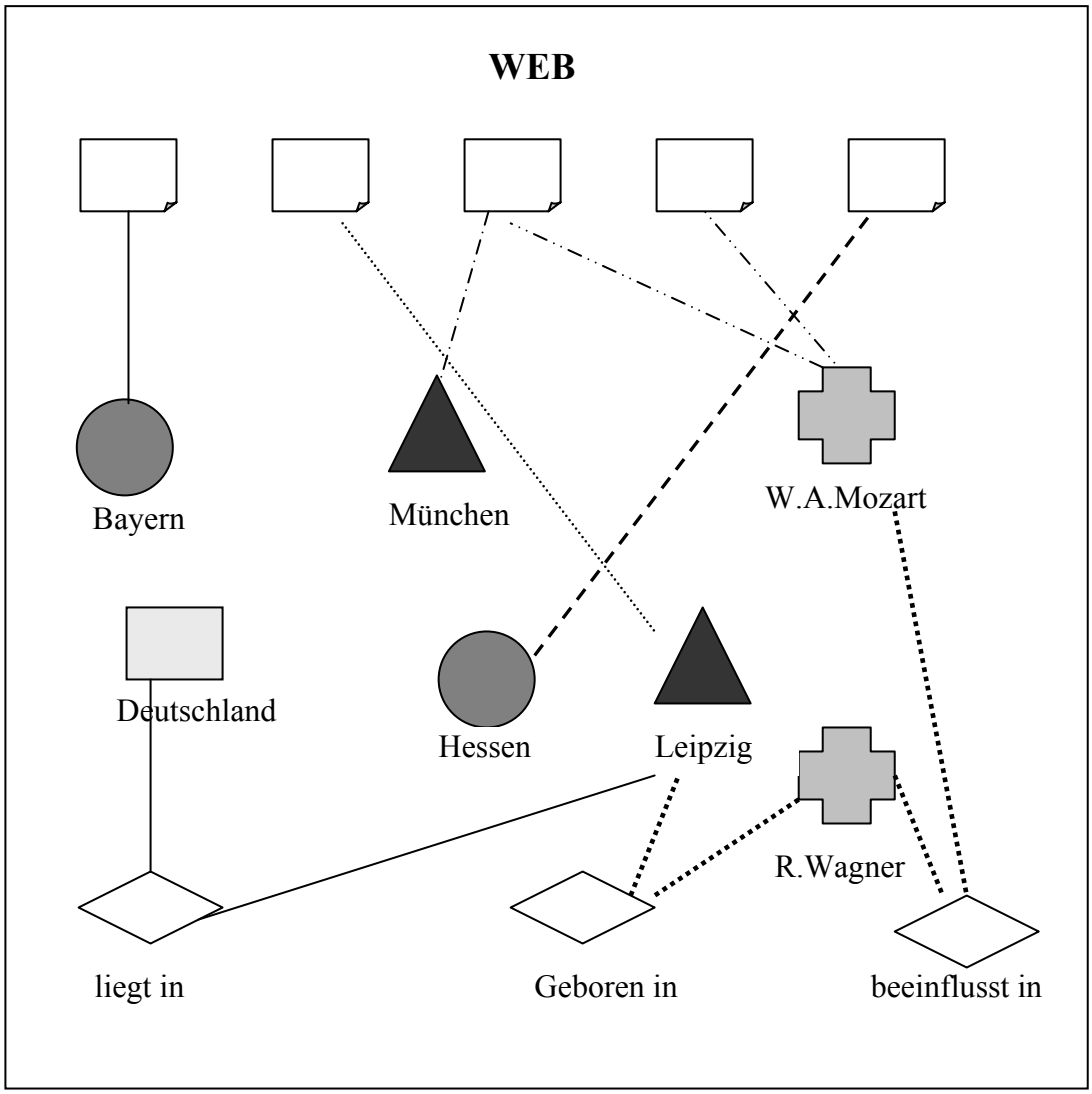


Abb.4 17: associations

4.5.4.7 Scopes

Um so größer eine Topic Map wird, desto größer ist auch die Wahrscheinlichkeit, dass darin Topics vorkommen, die den selben Namen habe, aber eine unterschiedliche Identität (vgl. Public Subjekt Descriptor). *Bayern* könnte einerseits der Name eines Bundeslandes und andererseits der Name eines Münchner Fußballclubs sein. Zur Lösung dieses Unterscheidungsproblems bietet der ISO Standard das Konzept der Scopes (Gültigkeitsbereiche). Topics können zu Scopes zugeteilt werden und sind dort eindeutig. So könnte man beispielsweise die beiden Scopes *Geographie* und *Fußball* definieren und *Bayern* je nachdem zuordnen(siehe Abb.4.18). Durch dieses Attribut kann angegeben werden, für welche Bereiche ein Topic Gültigkeit besitzt. Diese Eigenschaft könnte nach Ansicht des Autors bei der Identifikation von

Wissensstrukturen sehr hilfreich sein. Durch die Verwendung dieser Funktionalität wird die Trefferanzahl mitunter drastisch reduziert.

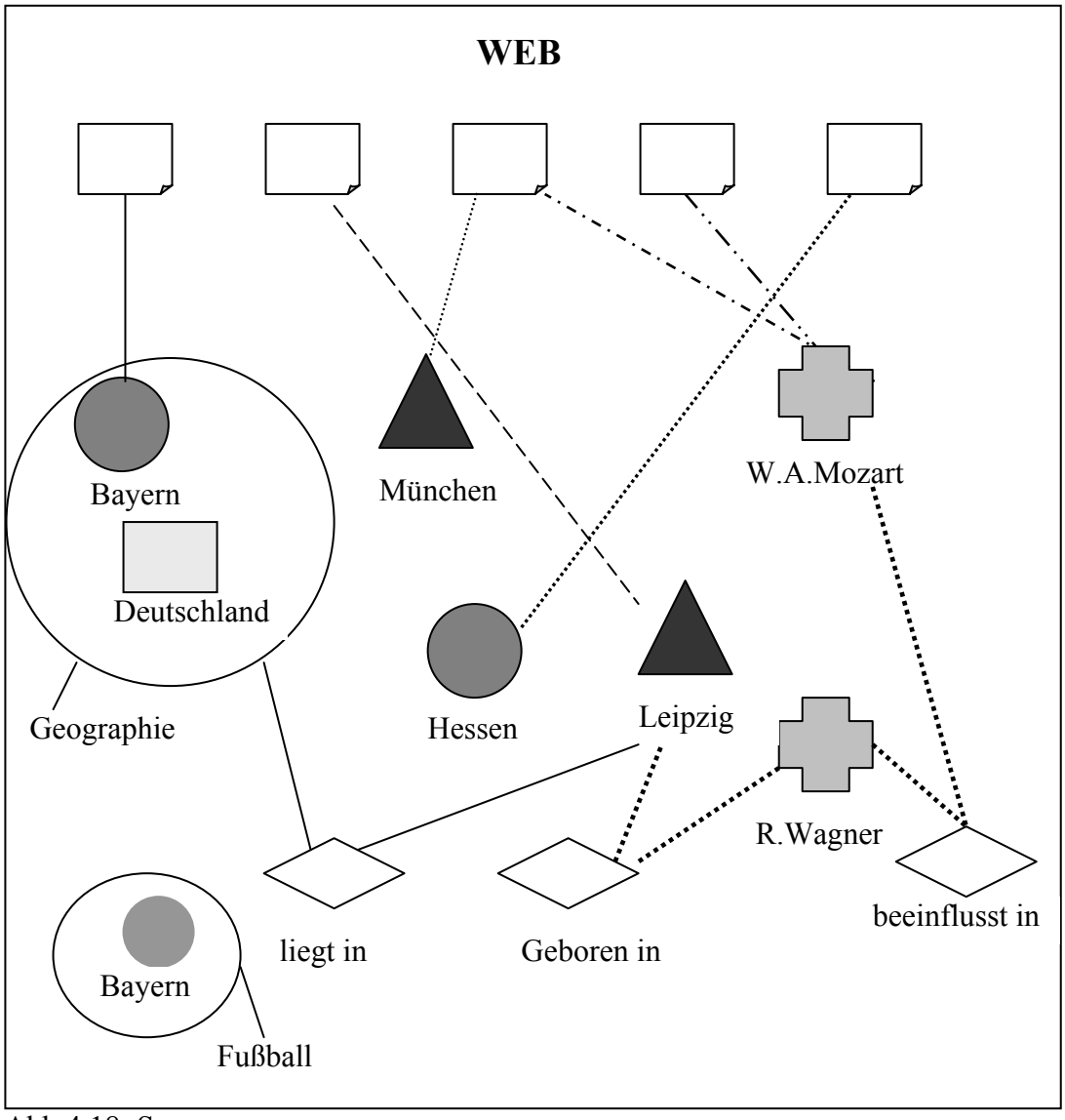


Abb.4.18: Scopes

4.5.4.8 Facets

Mit Facets (Facetten, Aspekte) können beliebigen Informationsquellen Eigenschaften bzw. Werte zugeordnet werden. Meist werden solche Facets einem Topic zugeteilt, jedoch können sie auch einer Association zugeordnet werden. Dem Topic München (in diesem Zusammenhang wieder als bayrische Landeshauptstadt) könnte etwa die Eigenschaft Einwohnerzahl mit dem Wert 1.300.000 beigefügt werden. Diese Facet könnte um eine weitere Facet, Jahr mit dem Wert 2001, erweitert werden. Dies bedeutet, dass München im Jahre 2001 1.300.000 Einwohner hat.. Es

besteht die Möglichkeit noch mehr Facets anzuhängen. Datentypen werden dabei nicht unterstützt. Man beschränkt sich hierbei auf Zeichenketten.
Die in den oben erläuterten Komponenten einer Topic Map können nun in Bezug auf unser Beispiel in XML dargestellt werden. Eine Darstellungsvariante in XML ist in Anhang B zu finden.

5 Abgrenzung verschiedener Methoden

5.1 KDD und OLAP

Mit Online Analytical Processing (OLAP) bezeichnet man die Analyse und Auswertung von multidimensional aufbereiteten Daten (Multidimensionale Daten), um Informationen für Unternehmensentscheidungen zu gewinnen. Die Informationen aus Rohdaten sind so destilliert worden, dass sie die realen Dimensionen eines Unternehmens aus der Benutzersicht wiedergeben. Schon bei einfachen Sachverhalten und Abfragen lässt sich dies erkennen, z. B. bei der Auswertung von Verkaufszahlen. Diese können nach unterschiedlichen Kriterien akkumuliert werden, wie nach Produkten (Produkteigenschaft, Produktkategorie, Industriezweig), Regionen (Stadt, Region, Land) oder der Zeit (Tag, Monat, Jahr). Zudem ist die Kombination einzelner Dimensionen denkbar, wie Produktkategorie, Stadt und Monat, Bei der Arbeit mit OLAP können neue oder unerwartete Beziehungen zwischen den einzelnen Variablen entdeckt und somit neue Anregungen für das Management abgeleitet werden.
KDD(*knowledge discovery in databases*) bezeichnet die Wissensgewinnung aus Datenbanken: nicht Daten oder Informationen sollen gewonnen werden, sondern echtes Wissen, das die Führungskräfte bei der Entscheidung Unterstützt. Wie OLAP benutzt KDD ein Data Warehouse als Datenbasis, geht aber bei der Auswertung der Daten über einfache Analyse-verfahren hinaus.
OLAP bezeichnet heute vor allem eine spezifische Analysefunktionalität, die über die überlicherweise von relationalen Datenbanksystemen angebotene Auswertungsfunktionalität hinausgeht. OLAP weist keinen Prozesscharakter auf. Im Gegensatz dazu ist KDD nicht einfach eine Sammlung von Verfahren, die auf Knopfdruck relevante Ergebnisse liefern, vielmehr ein komplexer Prozess, der zur Identifizierung von Mustern und Assoziationsregeln in der Datenbasis dient , um neue Erkenntnisse

mit hohem Nutzenpotential zu gewinnen. Aus Sicht des KDD-Prozesses spielen Methoden des OLAP eine Rolle als erweiterter Datenzugriff und Vorverarbeitung.

5.2 Data Mining und OLAP

Bei den eher traditionellen Methoden der Datenanalyse wie Reportgeneratoren aber auch bei OLAP-Systemen muss die zu prüfende Hypothese vom Anwender vorgegeben werden. Ein Experte wählt die geeignete Methode aus, mit der diese Hypothese durch Wechseln des Blickwinkels(Slicing, Dicing) oder der Aggregierungsgranularität (Drill-Down, Roll-Up) überprüft, verfeinert oder verworfen wird, und bereitet sie dann so auf, dass der Anwender wiederum die Ergebnisse interpretieren kann. Dabei besteht jedoch die Gefahr, dass man sich bei der Suche nach neuen Zusammenhängen auf eher subjektive Einschätzungen verlässt. Ein großes Problem stellt auch die Anfrageformulierung dar [Fay98b]: Will man beispielsweise in dem Datenbestand einer Bank diejenigen Datensätze filtern, die auf einen Kreditkartenbetrug hinweisen, so lässt sich eine solche Anfrage kaum in SQL formulieren.
Im Gegensatz hierzu verfolgt die Forschungsrichtung KDD dieselben Absichten selbständig, d.h. ohne zusätzliche Interaktionen mit dem Benutzer, der die Suche nach Auffälligkeiten nicht durch seine subjektiven Präferenzen beeinflussen und damit unbewusst die Suche einschränken soll. Dabei werden Hypothesen über mögliche Zusammenhänge, Muster oder Trends automatisch generiert und diese Hypothesen anhand von Daten überprüft. Aus der Vielzahl möglicher Hypothesen werden die gültigen als Ergebnis zurückgegeben. In Abb. 5.1 werden diese Unterschiede noch einmal verdeutlicht.

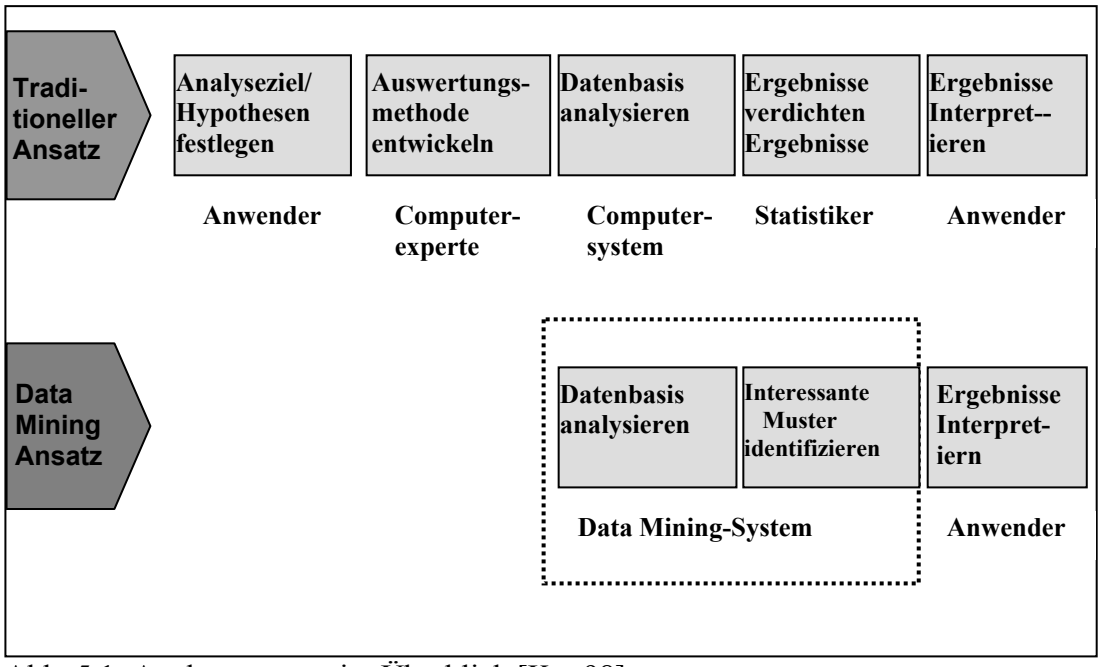


Abb. 5.1: Analyseansätze im Überblick [Küp98]

5.3 KDD und Data Mining

Auf der Basis der angeführten Begriffsdefinitionen (siehe im Abschnitt 4.1) wird der Begriff KDD als gesamten Prozess zur Wissensgewinnung und Wissensentdeckung aus einer großen Rohdatenmenge verstanden, der eine Reihe von Teilschritten beinhaltet. Der Begriff Data Mining beschreibt einen dieser Teilschritt, in dem die eigentliche Analyse der Daten vorgenommen wird. Die Vertreter der KDD-Anhänger verstehen unter Data Mining lediglich den Vorgang der Methodenanwendung auf den Datenbestand und integrieren Data Mining in einen übergreifenden KDD-Prozeß. Während einer typischen KDD-Anwendung werden eine Vielzahl von Hypothesen und (Zwischen-)Ergebnisse auf unterschiedlichen Datensätzen und mit unterschiedlichen Methoden erzeugt. In diesem komplexen und vielschichtigen Prozess braucht der Benutzer an verschiedenen Stellen Unterstützung. Unterstützung ist erforderlich bei der Bestimmung der Data Mining Ziele, bei der Auswahl der geeigneten Data Mining Methoden, bei der Interpretation der (Zwischen-)Ergebnisse, bei der Auswahl und Spezifikation nachfolgender Experimente und bei der Dokumentation des Prozesses und der Ergebnisse. Im Gegensatz dazu sucht und generiert das Data Mining aus den jeweiligen Daten neuartige Hypothesen nach vorgegebenen Mustern und bewertet diese autonom und/oder in Interaktion mit dem Experten.

Eine wiederholte, verfeinerte Formulierung und Validierung präzisiert mehr und mehr die Hypothese, die schließlich zu einem Modell der gefundenen Strukturen führt.

Da Data Mining Muster generiert, die Daten im Sinne von objektivem Wissen darstellen, intendiert das KDD-Konzept dagegen die Vermittlung von Information im Sinne eines subjektbezogenen Wissenszuwachses. Dem Data-Mining-Konzept fällt damit die Aufgabe eines formalen Mustergenerators zu, während das KDD-Konzept darüber hinaus Sorge zu tragen hat, dass die entdeckte Muster durch den Benutzer interpretiert werden können.

.

5.4 RDF und Topic Maps

Obwohl beide Konzepte semantische Netzwerke bilden können (Informationen vernetzen), weist RDF im Vergleich zu Topic Maps eine wesentlich einfachere Struktur auf. Aufgrund dessen ist es aber nicht in gleicher Weise so ausdrucksmächtig wie Topic Maps. Topic Maps ermöglichen Aufgrund ihres Aufbaus irrelevante Charakteristiken zu ignorieren; diese Möglichkeit bietet RDF nicht. RDF beschreibt zudem keine Zusammenhänge zwischen einzelnen Informationsressourcen. Dieses Problem wird erst durch RDF Schema gelöst. RDF Schema beschreibt grundlegende Typen von Properties. Damit ist es möglich, ähnlich mächtige Strukturen wie mit Topic Maps abzuleiten. Standardmäßig sind sogar Superklasse-Subklasse-Relationen vorgesehen. Über das Typ-Attribut kann ebenfalls eine Klasse-Instanz-Relation ausgedrückt werden. RDF-Schema bietet einen Standard an, mit dem man Ontologien abbilden kann, was bei Topic Maps im Moment noch nicht soweit ist. Ein weiterer Unterscheidungspunkt ist jener, dass Topic Maps themenorientiert und RDF ressourcenorientiert sind. Daraus stellt RDF keine echte Alternative zu Topic Maps dar. Gründe hierfür werden in der fehlenden Möglichkeit, Assoziationen zwischen verschiedenen Informationsressourcen herzustellen, gesehen. Aufgrund der primären Ausrichtung auf die Beschreibung von Informationsressourcen bietet es aber sehr gute Ergänzungsmöglichkeiten zum Topic Map Standard, da Topic Maps bei den Vorkommnissen(Occurrences) nur dürftige Beschreibungsmöglichkeiten bieten. Topic Maps stellen außerdem keine Konzepte wie Bags, Sequenzen oder Alternativen zur Verfügung. Durch die Einführung von Containerelementen zur Beschreibung von Reihenfolgebeziehungen könnte das

Einsatzpotential von Topic Maps im Wissensmanagement noch gesteigert werden. Die Entwicklergruppen haben bereits solche Synergieeffekte erkannt, und es bleibt abzuwarten, wie sich beide Konzepte weiterentwickeln. Die Tabelle 3 soll RDF und Topic Maps vergleichend gegenüberstellen:

Vergleichspunkt	RDF	Topic Maps
Syntax	XML	SGML, XML bzw. XTM
Hauptelemente	Description mit zugehörigem Property-Value-Paar	Topic, Association Occurrence
Zweck	Nähere Beschreibung von Informationsressourcen	Näherer Beschreibung von Dingen mit Ihren Zusammenhängen zu anderen Dingen,zusätzlich ist die Angabe von relevanten Informationsressourcen möglich
Besonderheiten	Containerelemente zur Beschreibung von Reihenfolgebeziehungen zwischen Informationsressourcen	Benennung der Topics(Topic Names) für verschiedene Zwecke möglich

Tabelle 3 : Vergleich RDF und Topic Maps[Hec01]

5.5 Information Mapping und Topic Maps

Information Mapping hat Ziel, Information so effektiv wie möglich zu strukturieren, um Navigation und Orientierung im Dokument zu erleichtern, Erfassen des Wesentlichen zu erhöhen und gezielt auf die gesuchten Information zuzugreifen. Topic Maps stellt das Ziel so dar: einfachste Navigation bzw. zielgerichtete Suche in großen, ständig wachsenden Informationsmengen(Dokument, Person etc.). Aufgrund dessen ist Information mapping eine Darstellungsmethode für das Dokument, in dem sich das Wissen wiederspiegelt. Topic Maps ist vielmehr ein Navigationsmethode, nach der sowohl die im Dokument enthaltenen Informationen als auch Dokument selbst gezielt gesucht werden können.

Obwohl Hyperlinks in Map- und Blocktiteln gemappte Informationen in jeder Datenbank gezielt und zuverlässig wiederauffindbar machen, weist Information mapping im Vergleich zum Topic Maps geringe Assoziationen zwischen Informationseinheiten auf. Im Gegensatz dazu beschreibt Topic Maps mithilfe von Document Type Definition (DTD), welche sämtliche möglichen Elemente und deren logische Reihenfolge untereinander definieren, Informationen über den Abschnitt selbst, aber auch über seine Beziehung zu anderen Elementen im Dokument. Diese Informationen können für die Verwaltung der Dokumente benutzt werden.

6 Einsatzmöglichkeiten

6.1 Data Mining

Die Menge möglicher Anwendungsfelder für das Data Mining ist ebenso unübersichtlich groß wie die Anzahl existierender Werkzeuge. Grundsätzlich sind Techniken der Wissensentdeckung immer dann einsetzbar, wenn ausreichend Daten vorhanden sind und eine Notwendigkeit der Analyse dieser Daten besteht[Nak98]. Daraus ist es hier auf einige wenige Beispiele beschränkt, bei denen die aktuelle Forschungsaktivitäten aufweisen könnten.

6.1.1 Geschäftliche Transaktionen

Der Geschäftsverkehr tendiert immer mehr dazu, sich in einer Weise zu konsolidieren, als Produkte bzw. Dienstleistungen zunehmend komplexer und damit spezialisierter werden. Während Unternehmen noch vor wenigen Jahre eine im Vergleich zu heute hohe Fertigungstiefe aufwiesen, werden aktuelle Entwicklungen von im wesentlichen zwei globalen Trends beschrieben: Globalisierung, einhergehend mit fortschreitender Unternehmensgröße, sowie Spezialisierung, einhergehend mit Ausgliederung von allen Arbeitsschritten, die nicht zu den jeweiligen Kernkompetenzen gerechnet werden. Verbunden damit ist eine Zunahme der Anzahl der Kunden eines Unternehmens, sowie der Anzahl der mit diesen abgewickelten Transaktionen. Man benötigt deswegen Abschätzungen für die so neu entstandenen Chancen und Risiken. Zu den sich in diesem Zusammenhang stellenden Fragen zählen:

- Wird dieser Kunde seine Rechnung bezahlen?
- Wird diese Transaktion mit betrügerischer Absicht durchgeführt?

- Welcher Erlös ist mit diesem Kunden in dem zurückliegenden Zeitraum verbunden gewesen, welcher Erlös ist in der Zukunft zu erwarten?
- Welche Produkte wird dieser Kunden als nächstes kaufen?

6.1.2 E-Commerce

Durch die Zunahme der Umstellung von geschäftlichen Transaktionen auf elektronische Verfahren sind zusätzlich zu den o. g. Anforderungen zumindest zwei weitere hinzugekommen: einerseits hat die Datenbasis an Umfang um mehrere Größenordnungen zugenommen, was die Analyse bzw. das auffinden von Mustern nicht gerade erleichtert; andererseits wird es zunehmend erforderlich, die Analyse in Echtzeit oder zumindest sehr zeitnah zu bewältigen

6.1.3 Webserver und Internetdaten

Einhergehend mit dem explosionsartigen Anstieg sowohl der Benutzer des Internets als auch der daran angeschlossenen Rechner gestaltet sich das Auffinden von Informationen zunehmend schwieriger. Suchmaschinen liefern in der Regel viele tausend Treffen für jeden gesuchten Begriff. Selbst die Kombination von Begriffen hilft Informationssuchenden oft nicht weiter. In dem Maße, wie die Informationsquelle im Internet effizient und effektive genutzt wird, entstehen daraus die große Herausforderung für die Data Mining Entwickler,

6.2 OLAP

Ein derzeit oft in der Literatur erwähnte Einsatzgebiet von OLAP ist Data Warehouses. *Data Warehouses* integrieren Datenbestände verschiedener (verteilter) operativer Datenbanken großer Unternehmen. Hier werden Informationen zusammengefasst, die den jeweiligen Entscheidungsträgern als Grundlage ihrer Unternehmenspolititk dienen sollen. Möglichst einfach zu bedienende OLAP (Online Analytical Processing) Systeme sollen eine komfortable Navigation durch diese multidimensionalen Datenbestände (Maßzählen wie z.B. Umsatz oder Gewinn, klassifiziert nach verschiedenen Dimensionen, wie Verkaufsgebiet, Zeitraum, Produkt) auf verschiedenen Aggregierungsebenen ermöglichen.

6.3 RDF

Dadurch dass RDF-Schemata je nach Bedarf selbst kreiert werden können, ergibt sich eine breite Palette von Anwendungsgebieten. Im folgenden werden einige genannt [Krm00]:

• *Klassifizierung von Internetseiten* nach formalbibliographischen Methoden wie in Bibliotheken: Durch die Katalogisierung von Internetseiten soll die Effizienz von Suchmaschinen gesteigert werden. Ein Beispiel hierfür ist das Open Directory Project (zu finden unter http://dmoz.org/)

• *Erstellen von Web-Sitemaps*: in den Metadaten können Angaben darübergemacht werden, wie einzelne Internetseiten einer Website zusammenhängen, sodass die Navigation erleichtert wird
• *Inhaltsbeschreibung von Internetseiten*
• *Bewertung von Internetseiten*: bietet in Kombination mit Filtersoftware die Möglichkeit, Seiten mit unerwünschten Inhalten auszuschließen. Als Beispiel kann hier PICS(Platform for Internet Content Selection), die einen Standard zur Erstellung von Metainformationen zu Dokumenten anbieten, genannt werden
• „*Privacy Practice Descriptions*“: bieten Informationen darüber, wie eine Seite mit persönlichen Informationen umgeht
• *Angaben zu digitalen Signaturen*: z.B. Informationen darüber, wie lange die Unterschrift noch gültig ist
• *vCard*: Bei der vCard handelt es sich um elektronische Visitenkarten, die vom Internet Mail Consortium entwickelt wurden. Die elektronische Form der Visitenkarte soll die Nachteile der Papier-Visitenkarten (z. B. Platzmangel) aufheben und können beispielsweise an Mails angehängt werden.

6.4 Topic Maps

6.4.1 Dokumentenmanagement

In Dokumentenmanagementsystemen (DMS) werden elektronische Dokumente durch die Vergabe von Indizes und Schlagwörtern näher beschrieben und im System abgelegt. Auch nicht in elektronischer Form vorliegende Dokumente können durch

die Verwendung von Scannern im System gespeichert werden. Mit Hilfe von Suchanfragen mit entsprechenden Angaben der Schlagwörter werden die Dokumente wieder auffindbar gemacht. Diese Vorgehensweise birgt aber auch Nachteile in sich. So ist zum Beispiel zum Auffinden eines bestimmten Dokumentes die Kenntnis der Schlagwörter notwendig. Wenn das System aber nur eine bestimmte Anzahl dieser Schlagwörter zulässt, so erhöht sich die Wahrscheinlichkeit, dass mehrere Dokumente mit der gleichen Beschreibung ausgestattet sind und somit wächst bei einer Suche mit den angegebenen Schlagwörtern auch die Trefferanzahl. Ein weiterer Nachteil liegt in den Dokumenten selbst. Jedes Dokument enthält nahezu immer mehr Informationen als für den eigentlichen Verwendungszweck prinzipiell benötigt werden. So enthält eine Rechnung beispielsweise nicht nur den zu zahlenden Endbetrag (der eigentliche Verwendungszweck) sondern noch die zugehörige Auftragsnummer, den Bearbeiter, das Erstelldatum, die Rechnungspositionen usw. Diese zusätzlichen Informationen sind notwendig, um Beziehungen zu zugehörigen Dokumenten herstellen zu können. Zusätzliche Schwierigkeiten bereitet die Extraktion von Informationen, die in Fließtexten „untergebracht“ sind. Alle diese Informationen sind im Dokument selbst „versteckt“ und die referenzierten Dokumente werden oft nur durch weitere Suchläufe gefunden. Der Nachteil liegt also in der „losen“ Speicherung einzelner Dokumente ohne erkennbaren Zusammenhang. Besonders deutlich wird dieser Nachteil bei der Verwendung eingescannter Dokumente, aus denen sich diese Informationen entweder gar nicht oder nur mit hohem Aufwand für eine maschinelle Verwertung gewinnen lassen. Des weiteren ist es erforderlich, die im Dokument enthaltenen Informationen für einen Leser schnell erkenntlich zu machen.

Die Lösung für den Einsatz von Dokumentbausteinen wird im Einsatz von einer inhalts-orientierten Dokumentauszeichnung auf Basis von SGML/XML gesehen. Dabei werden den unterschiedlichen Dokumentarten spezifische Strukturen zugeordnet, die für alle Dokumente dieser Art gelten. Eine solche Strukturbeschreibung heißt DTD - Document Type Definition-. Der Inhalt eines jeden Dokumentes wird dann dem entsprechenden Feld der Struktur zugeordnet. Über entsprechende Software kann dann auch in großen Datenbeständen inhaltsorientiert gesucht werden.

Mit dem Prinzip der Topic Maps sind Gemeinsamkeiten zur Speicherung von Dokumenten in herkömmlichen Dokumentenmanagementsystemen auf den ersten Blick leicht feststellbar. Es wird beiderseits ein Informationsressourcenpool

aufgebaut, der durch einzelne Schlagwörter referenziert wird. Topic Maps bieten zusätzlich weitere Möglichkeiten. Es ist jetzt auf einfache Art und Weise möglich, Zusammenhänge zwischen einzelnen Dokumenten zu modellieren. Dabei wird jedes Dokument durch ein Topic mit dem zugehörigen Verweis auf die Informationsressource repräsentiert. Danach wird eine entsprechende Assoziation zwischen diesen Topics gebildet.

6.4.2 Internet

Aufgrund des exponentiellen Wachstums der permanent am Internet angeschlossenen Rechner steigt auch die Anzahl der zur Verfügung stehenden Dokumente im Internet an. Erschwerend wirkt dabei die Tatsache bei dem Suchen nach den relevanten Informationen, das es sich hierbei um einen völlig unstrukturierten und heterogenen Datenbestand handelt. Zusätzlich unterliegt eine große Anzahl an Dokumenten ständigen Veränderungen und Aktualisierungen. Einen Beitrag zur Lösung des oben genannten Problems sollen Suchmaschinen teilweise leisten. Der große Vorteil von Suchmaschinen liegt in der vollautomatischen Aktualisierung ihres Datenbestandes. Eingriffe von menschlicher Seite sind nur bei der Anmeldung neuer Seiten oder Domains nötig. Nachteilig wirkt sich die oftmals hohe Trefferquote bei geringer Relevanz aus. Als eine Ursache ist hierfür die Volltextindizierung und ein ungenügendes Ranking-System zu sehen. Kommt in einem Text ein Wort sehr häufig vor, so wird vom System her eine hohe Relevanz angenommen.
Es ist prinzipiell vorstellbar, dass alle indizierten Seiten durch eine Suchmaschine in Form einer Topic Map abgespeichert werden. Dazu müsste zu jedem Wort im Lexikon ein entsprechendes Topic angelegt und mit den Einträgen aus dem Repository mit den Vorkommen (Occurrences) versehen werden. Diese Schritte werden realisierbar sein. Problematisch ist es die Tatsache, dass damit die Potentiale von Topic Maps bei weitem noch nicht ausgeschöpft werden. Bei der Angabe der Vorkommen eines Topics sollte auf alle Fälle auch der Typ des Vorkommens (z.B. Biographie, Abbildung, Konferenztext,...) mit angegeben werden. Um diese Aufgabe automatisch vom Rechner durchführen zu lassen, müssten Verfahren zur Inhalts- und Sinnerkennung der bearbeiteten Texte zur Verfügung stehen.

7 Zusammenfassung und Ausblick

KDD ist derzeit sowohl in der Wissenschaft als auch in der kommerziellen Welt ein viel diskutiertes Thema. Auf der einen Seite sorgt die Interdisziplinarität des Gebietes für einen fruchtbaren Gedankenaustausch der unterschiedlichen Fachrichtungen, auf der anderen Seite führen die vielfältigen Anforderungen aus der Praxis dazu, dass die Forschung nicht in akademischen Feinheiten stecken bleibt, sondern sich an realen Problemen orientiert und von diesen vorangetrieben wird.
Obwohl viele Fortschritt in KDD gemacht wurden, gilt es nach wie vor, eine Fülle von Herausforderungen zu meistern.

- Datenmengen im Bereich Giga und Terabyte sind heutzutage nicht selten. Die meisten eingesetzte Methoden und Systeme sind nicht oder nur mit Mühe in der Lage, direkt auf diesen Datenmengen zu lernen. So sind weitere Anstrengungen notwendig, um durch eine geschickte Verwaltung der Daten den KDD-Prozess auch auf großen Datenmengen effizient und effektive durchführen zu können.
- Bei der Bewertung und Interpretation von KDD-Ergebnissen müssen die Kriterien Nützlichkeit, Gültigkeit, Verständlichkeit, Interessantheit und Unerwartetheit berücksichtigt werden.
- KDD ist interaktiv, d.h. der Benutzer ist in den Prozess eingebunden. Allerdings kann er derzeit nur in begrenztem Umfang sein Vorwissen und seine Erwartungen explizit der Maschine mitteilen.

Diese Herausforderungen werden derzeit in KDD bearbeitet. Auch wenn noch viele Fragen ungelöst bleiben, steigt die Zahl der Firmen und Organisationen , die mit Hilfe von KDD die Flut der angehäuften Daten in nutzbringende Information verwandeln wollen, rapide und stetig an.

Data Mining basiert im wesentlichen auf Fortschritten der letzten zwanzig Jahre auf Gebieten einerseits wie technischen Fortschritten auf den Gebieten der Computer- und Netzwerkentwicklung andererseits. So zeigt sich Data Mining immer deutlicher als neue Disziplin im bedeutenden Anwendungen in Forschung und Entwicklung, Gesundheitsfürsorge, Bildung und Geschäftsleben. Ohne Bezugsnahme auf konkrete Untersuchungsverfahren kann Data Mining als datengetriebene Form der Daten-

analyse angesehen werden. Erst durch konsequente Nutzung dieses Ansatzes und Kombination mit hypothesengetriebenen Untersuchungsansätzen lassen sich die Potenziale, die sich durch die Analyse großer betriebswirtschaftlicher Datenbestände ergeben, richtig ausschöpfen. Datenanalysen führen nur durch organisatorische Einbettung in einen übergeordneten Prozess der Wissensentdeckung zu nutzbaren Ergebnissen, der durch erheblich Komplexität gekennzeichnet ist und neben der eigentlichen Analyse eine Reihe weiterer Aufgaben umfasst. Data Mining wird vorangetrieben durch das explosionsartige Wachstum komplexer Datenbestände sowie den relativen Mangel qualifizierter Wissenschaftler und Ingenieure, die in der Lage sind, solche Datenvolumina zu analysieren. Data Mining beginnt eigene Forschungs- und Entwicklungsaktivitäten zu entfalten, wobei Themen wie Assoziation, Visualisierung, Miningverfahren und Algorithmen für die Erkundung komplexer und verteilter Datensätze im Vordergrund stehen.
Trotz aller bisherigen Entwicklung steht das Data Mining noch am Anfang der Forschung. Es existieren immer noch viele ungelöste Probleme, wie z.B. eine einheitliche und flexible Sprache für die Datenmustererkennung oder Data Mining in Internetinformationssystemen. Deshalb kann auf eine Weiterentwicklung der vorhandenen Systeme nicht verzichtet werden. Insgesamt jedoch verspricht diese Technik den Umgang mit den immer größeren Datenmengen zu verbessern und somit die Ergebnisse für das Gewerbe effektiv ausnutzen zu können.

OLAP ist eine sehr leistungsfähige Architektur, die sich für die heutigen Anforderungen an die Verfügbarkeit von Daten in Data Warehouses gut anpasst. Die Einbindung multidimensionaler Datenbestände in die Domäne von WWW-Seite kann als sinnvoller Schritt angesehen werden, um den Verteilungsaspekt von unternehmensweiten Planungsdaten und Berichten im Umfeld von Intranet-Lösungen zu bewältigen. aber auch die Anreicherung der OLAP-Oberflächen um multimediale Darstellungsformen ist wünschenswert. In der absehbaren Zeit wird eine Reihe von OLAP-Anwendungen , sowohl im kommerziellen Bereich als auch in Form von Individualentwicklungen, auftauchen. Dann werden auch die Schwächen der Architektur zutage treten, die momentan noch nicht vorstellbar sind.

RDF ist eine Empfehlung des World Wide Web Consortiums. In ihm wird ein Datenmodell, eine mögliche Syntax und ein flexibles Klassensystem für Metadaten

über verschiedenste Arten von Ressourcen, insbesondere Web-Ressourcen, definiert. Ziel ist dabei, die Metadaten nicht wie bisher nur maschinenlesbar bereitzustellen, sondern mittels der in den Schemas des Klassensystems enthaltenen Semantik auch maschinenverständlich zu machen. RDF stellt damit ein flexibles und ausdrucksmächtiges System zur anwendungsübergreifenden Repräsentation von Metadaten zur Verfügung.

"The Resource Description Framework (RDF) is used in Navigator 5 for many purposes. Its aswiss army knife and we will use it wherever it makes sense to use the rdf data model as a representation language." [Chur]

Diese Aussage macht deutlich, dass das RDF nicht nur akademisch interessant ist, sondern sofort in der Praxis benutzt wird. Bedarf ist vorhanden, da die Dokumentenverwaltung bei der bisherigen Entwicklung des WWW insbesondere der Sprache HTML praktisch keine Rolle gespielt hat. Jetzt wird unter Führung des World Wide Web Consortium die Arbeit mit Dokumenten im Web auf eine neue zukunftsträchtige Grundlage gestellt. Im Mittelpunkt stehen dabei die eXtensible Markup Language (XML) und das Document Object Model (DOM), eine Plattform und sprachneutrale API zur Manipulation von Dokumenten. In diesen Kontext gehören neben RDF und RDF-Schemata auch die eXtensible Stylesheet Language (XSL), Namespaces in XML, XLink und XML-Data. Damit soll die Basis geschaffen werden, um das WWW zukünftig als zentrales Publikationsmedium und effizienten Informationspool nutzen zu können. Die neuen Standards sollen Einzellösungen von Firmen zuvorkommen. Sie bieten mit ihrer Flexibilität und Erweiterbarkeit hoffentlich genügend Potenzen, um auf eine breite Akzeptanz bei den Informationsanbietern zu stoßen.

Informationen, die so aufgeschrieben wurden, dass sie unverständlich oder auch nur schwer verständlich sind, sind nutzlos. Auch die besten technischen Realisierungen, diese Informationen allen Mitarbeitern eines Unternehmens zur Verfügung zu stellen, helfen da nichts. Dokumente, die nach festen Regeln strukturiert wurden – also z.B. Dokumente, die nach Information Mapping aufgebaut wurden – haben neben der erhöhten Verständlichkeit für den Benutzer noch einen weiteren entscheidenden Vorteil: Die Strukturen bergen Informationen, die durch den Einsatz entsprechender Tools ausgewertet und zum Wissen verwandelt werden können und den Nutzen der Dokumente noch steigern. Deshalb ist die Information-Mapping-Methode sehr

empfehlenswert, um ein neues Dokument zu erstellen. Die Frage, „Wie soll ich nun beginnen?“, erübrigt sich, da zur Erstellung ein klarer Ablauf vorgegeben ist. Das Qualitäts-Handbuch ist schon vorhanden. Es wäre eine vollständige Überarbeitung erforderlich, um die Information-Mapping-Methode auf eine bereite Basis anzuwenden.

Topic Maps sind die Lösung zur Organisation und Navigation der immer größer werdenden Informationsmenge. Der ISO-Standard liefert einen begrenzten, aber vollständigen und implementierbaren Satz an Konzepten und die beliebige Kombinierbarkeit dieser Konzepte ermöglicht eine große Vielfalt an Anwendungen, deren Grenzen zur Zeit eigentlich noch nicht abgeschätzt werden können. Zwei Anwendungsbereiche sind aber bereits schon absehbar: eine intelligente Erschließung des Webs und eine strukturierte Darstellung von Wissen. Topic Maps machen Ideen/Vorstellungen auch adressierbar. Sie werden eine deutliche Rolle spielen als eine Infrastruktur für das Semantic Web. Sie sind ein Mittelweg zwischen Dokumentation und Künstlicher Intelligenz(KI), eignen sich daher für das Wissensmanagement. Durch die semantische Hospitabilität ist der Ausbau in Richtung KI möglich.
Mit Topic Maps liegt erstmals eine standardisierte Möglichkeit vor, beliebige semantische Auszeichnungssprachen für begriffliche Wissensstrukturen auszutauschen und aufeinander abzubilden. Werden Grundlagen und Erfahrungen der Wissensorganisation beim Aufbau solcher Strukturen beachtet, kann sich mit Topic Maps ein Hebeleffekt einstellen, der die Bedeutung hochqualitativer Erschließungen (und die Rolle der Indexierer) unterstreicht. Es wird erwartet, dass die qualitative Vorteile von Topic Map bald in größeren Anwendungen offensichtlicher werden. durchaus kritische Gesichtspunkte. So scheint etwa das Verhältnis von XTM und dem Resource Description Framework (RDF) im Hinblick auf Konkurrenz oder Ergänzung noch völlig offen.

8 Literaturverzeichnis

[Alp00] Alpar, Paul; Niedereichholz, Joachim: Data Mining im praktischen Einsatz: Verfahren und Anwendungsfälle für Marketing, Vertrieb, Controlling und Kundenunterstützung; Fried. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig; 2000

[Agr93] Agrawal, R.; Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases; in *Proceedings of the ACM SIGMOD Conference on Management of Data*; Washington D.C.; Mai 1993; Seite 207-216

[Bac96] Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; Weiber, Rolf: Multivariate Analysemethoden: Eine anwendungsorientierte Einführung; 8. Auflage; Springer-Verlag; Berlin; 1996

[Ben99] Bensberg ,Frank: Das Data-Mining-Konzept, Juni 1999 http://www.wi.uni-muenster.de/aw/publikationen/CGC8.pdf

[Bie00] Michel Biezunski, Martin Bryan, Steven R. Newcomb, Topic Maps: Information Technology – Document Description and Markup Languages, 3.Dec 1999 http://www.y12.doe.gov/sgml/sc34/document/0129.pdf

[Bor98] Borgelt, Christian; Kruse, Rudolf: Attributauswahl für die Induktion von Entscheidungsbäumen: Ein Überblick; in *Data Mining – Theoretische Aspekte und Anwendungen*; Physica-Verlag; Heidelberg; 1998; Seite 77-98

[Bre98] Breitner,C.A.; Lockemann, P.C; Schlösser J.A. : Die Rolle der Informationsverwaltung im KDD-Prozess; in *Data Mining –Theoretische Aspekte und Anwendungen*; Physica-Verlag; Heidelberg; 1998: Seite 34-60

[Bug93] Buggle, Franz: Die Entwicklungspsychologie Jean Piagets. 2. Auflage Kohlhammer, Stuttgart 1993

[Cha99] Chamoni, Peter; Gluchowski, Peter: Analytische Iinformationssysteme: Data Ware- hause, On-Line Analytical Processing, Data Mining; Springer-Verlag; Berlin, 1999 Seite 262-280

[Chu73] Churchman, C.West: die Konstruktion von Erkenntnissystemen. Herd&Herd Frankfurt/New York, 1973

[Chur] Churchill, Guha, J. ;Giannandrea , R. :Netscape Communications Corporation; Mozilla.Org, http://www.mozilla.org/rdf/doc/index.html

[Düs98] Düsing, Roland: Knowledge Discovery in Databases und Data Mining; in *Analytische Informationssysteme*; Springer-Verlag; Berlin; 1998; Seite 291-299

[Fay96a] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy: Advances in Knowledge Discovery and Data Mining; AAAI Press, California; 1996

[Fay98b] Fayyad, Usama: Mining Databases: Towards Algorithmus for Knowledge Discovery; in Bulletin of the Technical Committee on Data Engineering; Vol. 21 No.1; März 1998; Seite 39-48

[For99] Forst, Annelyse: Information und Wissen(Teil 1): Die neuen betrieblichen Ressourcen; 1999
http://www.doculine.com/news/1999/Februar/infowiss.htm#1

[Fre00] Freese, E.: Using Topic Maps for the representation, management and discovery of knowledge; 2000
http://www.gca.org/papers/xmleurope2000/papers/s22-01.html

[Hec01] Heckel, R. : Einsatzmöglichkeiten von Topic Maps zur flexiblen Navigation in elektronischen Dokumenten, Diplomarbeit, Technische Universität Dresden; 2001

[Hol00] Holzmann, Martin; Information Mapping: Vorteile durch effizientes Strukturieren; 2000 Januar ttp://www.doculine.com/news/2000/Januar/infomap.htm

[Imm99] Die Information Mapping Methode 30 Jahre Forschung; Ein Überblick; 1999 http://www.carstens-techdok.de/TD/IMAGES/IMAP.PDF

[Kra98] Krahl, Daniela; Windheuser, Ulrich; Zick, Friedrich-Karl: Data Mining: Einsatz in der Praxis; Addison Wesley Longman Verlag, Bonn; 1998

[Krm00] Kratzert, M; Matthes, M. : Metadaten-Konventionen und PKM; Deutschland; 2000
www.cmr.fuberlin.de/~mck/courses/v00ss/PeKMan/team4/metadatenkapitel.pdf

[Küp98] Küppers, Bertram: Data Mining in der Praxis: Ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld; Europäische Hochschulschriften:Reihe 5; Volks- und Betriebswirtschaft Bd. 2373; Lang Verlag, Frankfurt am Main; 1998

[Las99] Lassila, O.; Swick, R. R.: Resource Description Framework (RDF) Model and Syntax Specification, World Wide Web Consortium,
http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[Mar75] Mary Ann S. , Pulaski: Piaget Eine Einführung in seine Theorie und sein Werk, 1975

[Nak98] Nakhaeizadeh, Gholamreza: Reinartz, Thomas; Wirth, Rüdiger: Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick; in *Data Mining –Theoretische Aspekte und Anwendungen*; Physica-Verlag; Heidelberg; 1998: Seite 3-29

[Pep99] Pepper, S., Navigating Haystacks, Discovering Needles, http://www.ontopia.net/topicmaps/materials/mlangart.pdf

[Pia91] Piaget, Jean: Meine Theorie der geistigen Entwicklung ; Frankfurt am Main Fischer-Taschenbuch-Verlage; 1983

[Pop99] Pop, Gunther: Einführung in OLAP; dc soft GmbH; 1999 http://www.dcsoft.de/knosys/html/body_olap.html

[RDF00], Resource Description Framework Schema Specification 1.0; 2000 http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

[RDFS99] Resource Description Framework; Model and Syntax Specification 1999 http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[Rth] Rath, Hans Holger : Mit Topic Maps intelligente Informationsnetze aufbauen; Mozart und Kugeln; http://empolis.com/englisch/pdf/Mozart-und-Kugeln.pdf

[Saü00] Saüberlich, Frank: KDD und Data Mining als Hilfsmittel zur Entscheidungsunterstü-rtzung. Europäischer Verlag der Wissenschaften Frankfurt am Main; 2000

[Smi01] Smolmik, Stefan: K-Discovery: Identifikation von verteilten Wissens-strukturen in einer prozessorientierte Groupware-Umgebung, in: Conference Proceedings ISKO 2001, Berlin, März, 2001

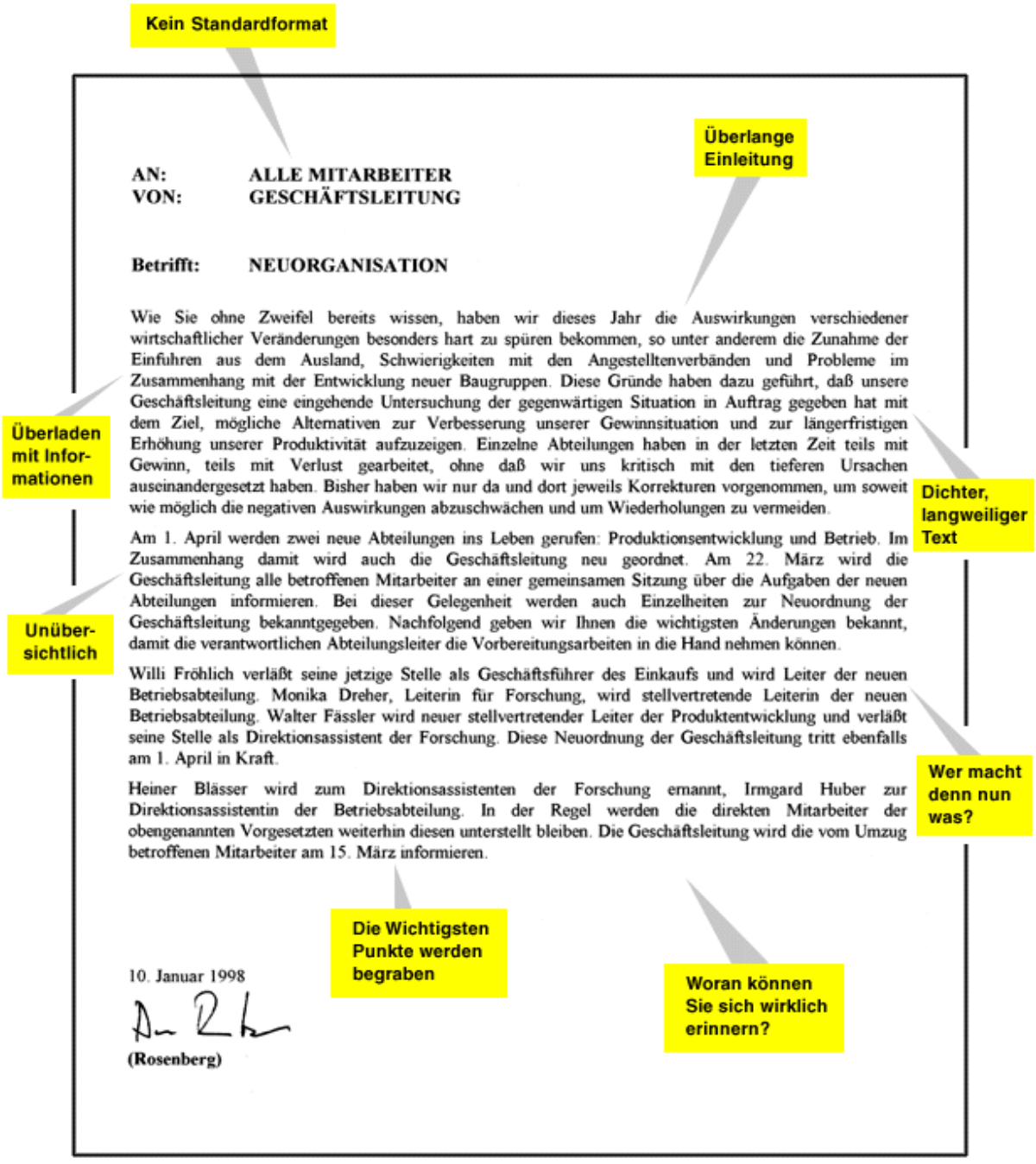
[Test] CARSTENS + PARTNER GmbH; Machen Sie den Information Mapping Test; http://www.carstens-techdok.de/TD/INDEX04B.HTM

9. Abkürzungsverzeichnis

CART	classification and regression trees
CHAID	Chi-squared automatic interaction detection
DC	Dublin Core
DMS	Dokumentenmanagementssystemen
DOM	Document Object Model
DTD	Document Type Definition
GPS	The Global Positioning System
HOLAP	Hybrides OLAP
IMP	Information Mapping
KI	Künstlicher Intelligenz
KDD	Knowledge Discovery in Database
MDBMS	Relational Database Management Systems
MOLAP	Multidimensional OLAP
OLAP	ON-Line analytical Processing
OLTP	On-Line Transactional Processing
RDBMS	Relational Database Management Systems
RDF	Resource Description Framework
ROI	Return of Investment
ROLAP	Relationales OLAP
SGML	Standard Generalized Markup Language
W3C	World Wide Web Consortiums
XML	Extensible Markup Language
XSL	eXtensible Stylesheet Language

Anhang A

Mit Normaltext



AN: ALLE MITARBEITER
VON: GESCHÄFTSLEITUNG

Betrifft: NEUORGANISATION

Wie Sie ohne Zweifel bereits wissen, haben wir dieses Jahr die Auswirkungen verschiedener wirtschaftlicher Veränderungen besonders hart zu spüren bekommen, so unter anderem die Zunahme der Einfuhren aus dem Ausland, Schwierigkeiten mit den Angestelltenverbänden und Probleme im Zusammenhang mit der Entwicklung neuer Baugruppen. Diese Gründe haben dazu geführt, daß unsere Geschäftsleitung eine eingehende Untersuchung der gegenwärtigen Situation in Auftrag gegeben hat mit dem Ziel, mögliche Alternativen zur Verbesserung unserer Gewinnsituation und zur längerfristigen Erhöhung unserer Produktivität aufzuzeigen. Einzelne Abteilungen haben in der letzten Zeit teils mit Gewinn, teils mit Verlust gearbeitet, ohne daß wir uns kritisch mit den tieferen Ursachen auseinandergesetzt haben. Bisher haben wir nur da und dort jeweils Korrekturen vorgenommen, um soweit wie möglich die negativen Auswirkungen abzuschwächen und um Wiederholungen zu vermeiden.

Am 1. April werden zwei neue Abteilungen ins Leben gerufen: Produktionsentwicklung und Betrieb. Im Zusammenhang damit wird auch die Geschäftsleitung neu geordnet. Am 22. März wird die Geschäftsleitung alle betroffenen Mitarbeiter an einer gemeinsamen Sitzung über die Aufgaben der neuen Abteilungen informieren. Bei dieser Gelegenheit werden auch Einzelheiten zur Neuordnung der Geschäftsleitung bekanntgegeben. Nachfolgend geben wir Ihnen die wichtigsten Änderungen bekannt, damit die verantwortlichen Abteilungsleiter die Vorbereitungsarbeiten in die Hand nehmen können.

Willi Fröhlich verläßt seine jetzige Stelle als Geschäftsführer des Einkaufs und wird Leiter der neuen Betriebsabteilung. Monika Dreher, Leiterin für Forschung, wird stellvertretende Leiterin der neuen Betriebsabteilung. Walter Fässler wird neuer stellvertretender Leiter der Produktentwicklung und verläßt seine Stelle als Direktionsassistent der Forschung. Diese Neuordnung der Geschäftsleitung tritt ebenfalls am 1. April in Kraft.

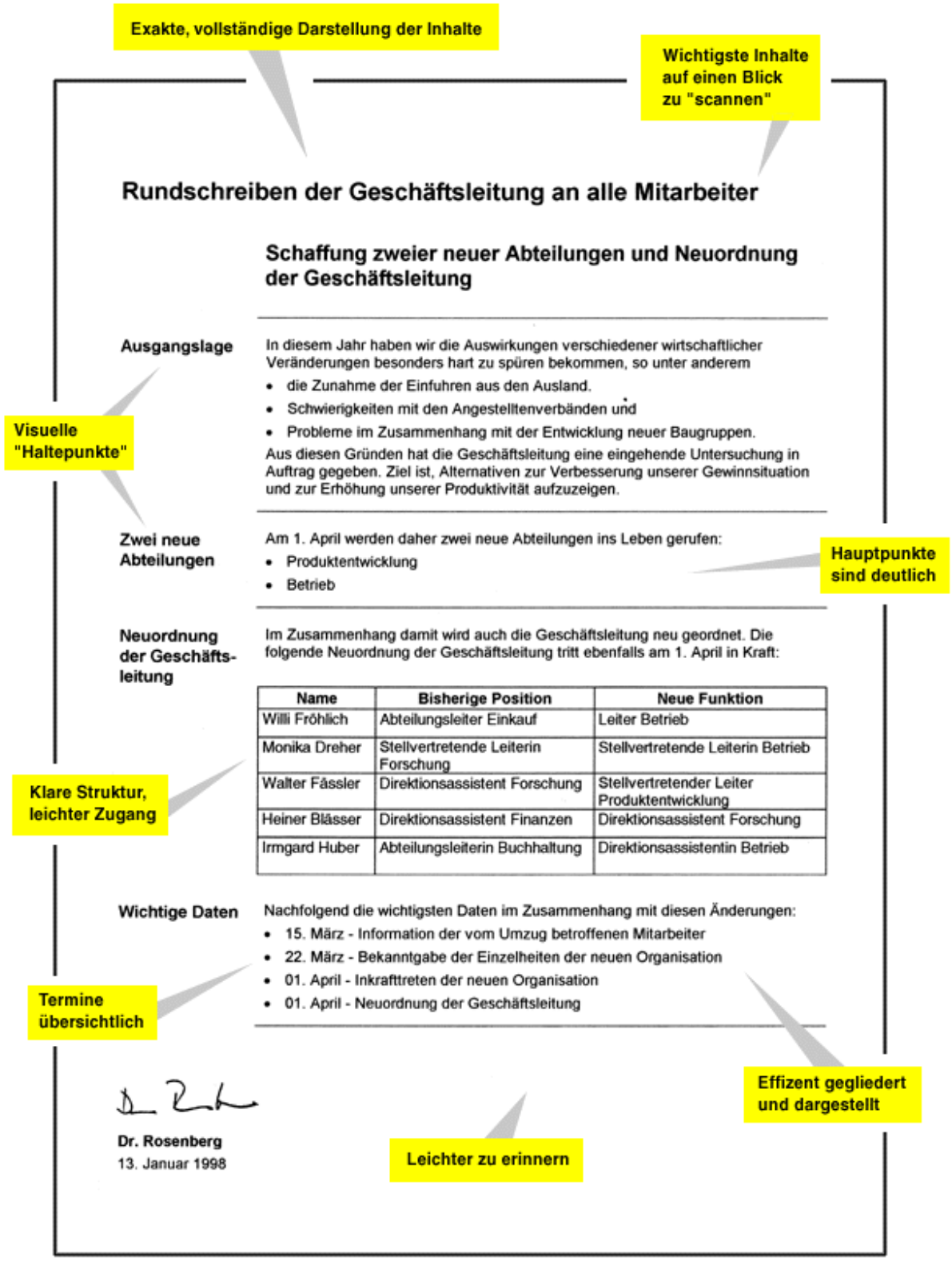
Heiner Blässer wird zum Direktionsassistenten der Forschung ernannt, Irmgard Huber zur Direktionsassistentin der Betriebsabteilung. In der Regel werden die direkten Mitarbeiter der obengenannten Vorgesetzten weiterhin diesen unterstellt bleiben. Die Geschäftsleitung wird die vom Umzug betroffenen Mitarbeiter am 15. März informieren.

10. Januar 1998

(Rosenberg)

Anhang A

Mit Information Mapping Text



Anhang B

```
<!--Hier werden die Topic Types mit ID und Topic Name deklariert-->
<topic id="tt-staat">
<topname><basename>Staat</basename></topname>
</topic>
<topic id="tt-bundesland">
<topname><basename>Bundesland</basename></topname>
</topic>
<topic id="tt-stadt">
<topname><basename>Stadt</basename></topname>
</topic>
<topic id="tt-person">
<topname><basename>Person</basename></topname>
</topic>

<!--Hier werden die Occurrence Role Types mit ID und Topic Name deklariert-->
<topic id="or-webpage">
<topname><basename>Webpage</basename></topname>
</topic>
<topic id="or-foto">
<topname><basename>Fotografie</basename></topname>
</topic>
<!--Hier werden die einzelnen Topics mit ID, den verschiedenen Topic Names und
den
zuvor deklarierten Topic Types und Occurrence Role Types deklariert-->
<topic id="t-deutschland" types="tt-staat">
<topname>
<basename>Bundesrepublik Deutschland</basename>
<dispname>BRD</disname>
<sortname>Deutschland</sortname>
</topname>
<occurs type="or-webpage">http://www.deutschland.de</occurs>
</topic>
```

<topic id="t-hessen" types="tt-bundesland">
<topname><basename>Hessen</basename></topname>
<occurs type="or-webpage">http://www.hessen.de</occurs>
</topic>
<topic id="t-bayern" types="tt-bundesland">
<topname><basename>Bayern</basename></topname>
<occurs type="or-webpage">http://www.bayern.de</occurs>
</topic>
<topic id="t-muenchen" types="tt-stadt">
<topname><basename>München</basename></topname>
<occurs type="or-webpage">http://www.munich.de</occurs>
</topic>
<topic id="t-leipzig" types="tt-stadt">
<topname><basename>Leipzig</basename></topname>
<occurs type="or-webpage">http://www.leipzig.de</occurs>
</topic>
<topic id="t-wagner" types="tt-person">
<topname>
<basename>Richard Wagner</basename>
<dispname>R. Wagner</disname>
<sortname>Wagner, Richard</sortname>
</topname>
<occurs type="or-foto">wagner.gif</occurs>
</topic>
<topic id="t-mozart" types="tt-person">
<topname>
<basename>Wolfgang Amadeus Mozart</basename>
<dispname>W. A. Mozart</disname>
<sortname>Mozart, Wolfgang Amadeus</sortname>
</topname>
<occurs type="or-foto">mozart.gif</occurs>
</topic>
<!--Hier werden die Association Types mit ID und Topic Name deklariert-->
<topic id="at-liegt-in">

<topname><basename>liegt in</basename></topname>
</topic>
<topic id="at-geboren-in">
<topname><basename>geboren in</basename></topname>
</topic>
<topic id="at-beeinflusst-von">
<topname><basename>beeinflusst von</basename></topname>
</topic>

Erklärung

Ich versichere hiermit, dass ich diese Arbeit selbstständig verfasst und nur die angegebenen Hilfsmittel verwendet habe.

.....
Unterschrift