Antrag auf Gewährung einer Sachbeihilfe

1 Allgemeine Angaben

Neuantrag

1.1 Antragsteller

Stefan **Conrad**, PD Dr. rer. nat. habil.
Dienststellung: wissenschaftlicher Assistent (C1)
Geburtsdatum: 11.08.1966
Nationalität: Deutsch
Dienstanschrift:
Institut für Technische u. Betriebliche Informationssysteme
Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2
39106 Magdeburg
Telefon: (++49) 391 67 18066
Telefax: (++49) 391 67 12020
E-Mail: `conrad@iti.cs.uni-magdeburg.de`

Myra **Spiliopoulou**, Dr. rer. nat. Griechenland
Dienststellung: Wissenschaftliche Assistentin (C1)
Geburtsdatum: 25.02.1965
Nationalität: Griechisch
Dienstanschrift:
Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Institut für Wirtschaftsinformatik
Spandauer Str. 1
10178 Berlin
Telefon: (++49) 30 2093-5716
Telefax: (++49) 30 2093-5741
E-Mail: `myra@wiwi.hu-berlin.de`

1.2 Thema

DIAsDEM: “**D**aten**i**ntegration von **A**ltlasten und **s**emistrukturierten **D**okumenten durch **E**insatz von **M**ining Verfahren”

1.3 Kennwort

DIAsDEM

1.4 Fachgebiet und Arbeitsrichtung

Fachgebiet:	Praktische Informatik
Arbeitsrichtung:	Datenbanken und Informationssysteme (Daten(bank)integration und Knowledge Discovery)

1.5 Voraussichtliche Gesamtdauer

Die voraussichtliche Gesamtdauer für das hier beantragte Projekt beträgt drei Jahre. Bei erfolgreichem Verlauf des Projektes ist eine Fortführung in einem Folge- bzw. Anschlußprojekt denkbar.

1.6 Antragszeitraum

Mittel werden für drei Jahre beantragt.

1.7 Gewünschter Beginn der Förderung

Gewünschter Beginn der Förderung ist der 1.07.1999.

1.8 Zusammenfassung

In der modernen Informationsgesellschaft erleben wir eine Überflutung an Informationen von potentieller Wichtigkeit und Nutzbarkeit. Zugleich wird die Nutzung von Informationen zunehmend selektiv, die Anforderungen an sinnvolle, verständliche und verwertbare Daten immer höher. Diese wertvollen Daten sind meistens nicht in direkt anwendbarer Form vorhanden, sondern liegen in veralteten Datenbanken oder als Bestände mit unklarer Struktur in oft sehr heterogenen Datensammlungen vor. Ziel von DIAsDEM ist daher die Integration von Altlastdatenbeständen und semistrukturierten Daten in einem Informationssystem, um sie zusammen mit konventionellen Daten befragen und nutzen zu können, z.B. zur Entscheidungsunterstützung in Unternehmen. Zur Erreichung dieses Ziels sollen Methoden des Data-Mining eingesetzt werden, mit folgenden Schwerpunkten: (i) Erkennung von semantischen Eigenschaften und Abhängigkeiten zwischen semantisch relevanten Daten und (ii) Nutzung der erkannten Eigenschaften zur Integration der Daten in einem Informationssystem. Bisher wurden solche Analysemethoden immer nur auf Daten aus einer einzelnen Quelle angewendet. In DIAsDEM werden wir sie für die Unterstützung der Integration mehrerer Datenbanken und -sammlungen einsetzen. In einem konkreten Anwendungsbereich aus der Bioinformatik sollen die in dem Projekt entwickelten Konzepte erprobt und validiert werden.

2 Stand der Forschung, Eigene Vorarbeiten

2.1 Stand der Forschung

DIAsDEM befaßt sich mit dem Problem der Koppelung von Altlastdaten und semistrukturierten Daten in einem Informationssystem. Wir streben eine Lösung zu der Entdeckung von Schemainformationen in den Daten und der Nutzung der so gewonnenen Metadaten für integrierte Anfragen an das Informationssystem an. Somit betrifft unser Forschungsvorhaben mehrere Teilgebiete, die hier kurz dargestellt werden.

Datenintegration

Die Probleme der Datenintegration werden im Bereich der föderierten Datenbanken ausführlich behandelt. Hier sind die Integration der Schemata und die nachfolgende Integration der Daten von Interesse.

Integration auf der Schemaebene. Neben den weitverbreiteten Schemaarchitekturen für föderierte Datenbanken [HM85, LMR90, SL90] gibt es zahlreiche Ansätze zur Schemaintegration. In [BLN86] werden bereits einige solcher Ansätze verglichen und allgemeine Anforderungen an Integrationsmethoden aufgestellt. Aktuellere Vergleiche, die neuere Ansätze berücksichtigen, sind unter anderem in [PBE95] (für objektorientierte Methoden) und in [Con97] zu finden.

Aufgrund der Vielzahl von Vorschlägen zur Schemaintegration seien hier nur wenige als Beispiele für konzeptionell unterschiedliche Ansätze genannt: In [SPD92] werden Zusicherungen (Korrespondenzen) zwischen Schemabestandteilen und spezielle Integrationsregeln als Basis für die Integration verwendet. Für die Integration objektorientierter Klassenhierarchien wird von vielen Ansätzen das *Upward Inheritance*–Prinzip [DH84, SN88] verwendet. Der in [Sch98] beschriebene Integrationsansatz führt eine spezielle Zwischenrepräsentation ein, die insbesondere die korrekte Behandlung extensionaler Zusammenhänge unterstützt.

Integration auf der Datenebene. Der Schemaintegration soll die Integration der Daten folgen, wobei es wichtig ist, zusammengehörige Daten als solche zu erkennen und möglichst exakt zuzuordnen. Schon für die klassischen Schemaintegrationsansätze lassen sich Fälle finden, in denen eine solche Zuordnung von Daten, die in verschiedenen Datenbanken dasselbe Real-Welt-Objekt repräsentieren, nicht eindeutig möglich ist (vgl. [Con97, S. 245ff]).

Die oft als Lösung vorgeschlagene Einführung einer `same`-Relation [HMS94, SS95, Che96] setzt exakte Kenntnisse des Experten und einen hohen Aufwand voraus. In der HIPED Architektur für Datenbankenintegration [NMO96] wird dieser Aufwand zum Teil behoben, indem ein deduktives Datenbanksystem Beziehungen zwischen Tabellen anhand von Beziehungen zwischen Attributen entdecken kann. Letztere müssen aber doch vom Experten mitgeteilt werden. Insbesondere für Altlastdatenbanken ist jedoch ein solcher Experte selten vorhanden. Außerdem ist der menschliche Arbeitsaufwand zur Entdeckung aller einzelnen zusammengehörigen Datenbestände aufgrund der Datenmengen in der Regel unannehmbar.

Einen Ausweg aus dieser Situation bietet die Tatsache, daß zusammengehörige relationale Daten meistens gewisse funktionale Abhängigkeiten („FA"s) oder Integritätsbedingungen

(„IB"s) erfüllen. Zwar sind in einer Altlastdatenbank FAs selten explizit beschrieben, sie können aber mit Mining-Verfahren von den Daten abgelesen und als Regeln ausgedrückt werden. Hier kommen Verfahren zur Entdeckung von Assoziationsregeln [ATS93], zur Klassifizierung [Dže96] und zur attribut-orientierten Induzierung [HF96] in Frage.

Solche Verfahren wurden jedoch konzipiert, um Beziehungen zwischen *Werten* statt zwischen *Attributen* zu finden. So können die Entdeckungen eines Mining-Verfahrens zu FAs generalisiert werden, die *nur mit einer gewissen Wahrscheinlichkeit gelten*. In [Rus98] wird informell ein Ansatz geschildert, der möglicherweise zur Gestaltung von probabilistischen FAs geeignet ist: Die Wertemenge jedes Attributs wird partitioniert, die Partitionen mit Wahrscheinlichkeitsintervallen versehen, und Regeln zwischen Partitionen hergeleitet.

Alle obige Studien beziehen sich auf Mining innerhalb *einer* relationalen Datenbank. Die verschiedenen Möglichkeiten zum Einsatz von Mining-Verfahren bei der Entdeckung, Transformation und Integration der Schemata von mehreren Datenbanken sind in [HC98] analysiert. Des weiteren zeigt die Studie von [SS96d], daß Mining auch zwischen Dokumentensammlungen möglich ist und sehr wichtige Einsichten über relevante Dokumente liefert.

In DIAsDEM betrachten wir die Entdeckung von FAs und IBs innerhalb jeder strukturierten oder semistrukturierten Datensammlung, sowie zwischen Datensammlungen, als einen Kernschritt zur Datenintegration [LNE89]. Dafür werden wir Mining-Verfahren verwenden und Lösungsansätze zu den oben geschilderten Einschränkungen entwickeln.

Modellierung von semistrukturierten Daten

Datenbestände außerhalb von Datenbanken unterliegen keinem klassischen Schema, wie es für die Datenintegration in einem Informationssystem erforderlich wäre. Bestrebungen zur Lösung dieses Problems führten zu Methoden für die Modellierung und Anfrage von semistrukturierten Daten. Darunter unterscheiden wir zwischen „schema-orientierten" und „schema-losen" Ansätzen.

Zu den schema-orientierten Ansätzen gehören (unter anderen) die Systeme von [AM98, AMM97, FFLS97], die eine vorgegebene Struktur der Daten annehmen. Abweichungen von dieser Struktur werden bestenfalls toleriert, indem die betroffenen Daten ignoriert werden. Dies bedeutet, daß Daten nur dann in einer Anfrage in Betrachtung kommen können, wenn ihre Struktur der vorgegebenen entspricht.

Zu den schemalosen Ansätzen gehört TSIMMIS [QRS+95, AQMW97], sowie das System von Buneman et al [BDHS96]. In beiden Fällen können Anfragen formuliert werden, in denen nach Unterstrukturen, Schlüsselwörtern oder Attributen in einer Datensammlung unbekannter Struktur gesucht wird. Somit kann eine Anfrage Datenbestände zurückliefern, die trotz unterschiedlicher Struktur die Anfragekriterien erfüllen.

In [Abi97] wird die Nutzung von Mining-Verfahren zur Schemaentdeckung in semistrukturierten Daten motiviert: *DataGuides* [GW97] werden dort als einsetzbares Werkzeug in dieser Richtung betrachtet. Für jede Datensammlung wird ein DataGuide entworfen, der die Struktur jedes Datenbestands abbildet. Der Entwurf eines DataGuide und seine Aktualisierung bei strukturellen Änderungen sind automatisiert. Allerdings ist dafür oft das Einlesen der ganzen Datensammlung erforderlich.

Data Mining zur Entdeckung von zueinander in Beziehung stehenden Dokumenten wird in [AFK97, FKZ97, PE98, LHKK96, ZEMK97] untersucht, ohne allerdings gemeinsame Eigenschaften zu extrahieren. In [CDAR97] werden sogenannte „Signatures" von Dokumenten entdeckt, um die Qualität der Treffer einer Textsuche im WWW zu verbessern.

In [WL97b, WL98] wird ein Mining-Verfahren zur Schemaentdeckung für semistrukturierte Daten vorgeschlagen, die laut dem OEM Graphenmodell [AQMW97] gestaltet sind: ein Unterbaum des Graphen wird zum „Schema-Muster" deklariert, wenn er eine maximale Größe aufweist und im Datengraph oft auftritt [WL97b].

Im allgemeinen wirkt die Musterentdeckung als das geeigneteste Paradigma für Schemaentdeckung in semistrukturierten Daten. Außer dem oben erwähnten Ansatz sind somit auch Verfahren wie [AS95, DCH95, MT96, Wan97, WL97a, SF98, Spi98] anwendbar, die sich in der Struktur der unterstützten Muster sowie in dem verwendeten Entdeckungsverfahren und dessen Annahmen unterscheiden.

Bei diesen Methoden zur Modellierung, Befragung und Strukturentdeckung bleibt aber noch die Frage unbeantwortet, wann zwei semistrukturierte Datenbestände dasselbe Real-Welt-Objekt darstellen oder in einer Abhängigkeitsbeziehung zueinander stehen. Wie oben geschildert, ist die Erkennung solcher Daten für die Integration auf der Datenebene erforderlich. So werden wir in DIAsDEM Mining-Verfahren einsetzen, um Strukturen und Inhalte zu entdecken, die auf identische oder zusammengehörige Dateninstanzen hinweisen.

Anfrageunterstützung nach der Datenintegration

Die Ergebnisse eines Mining-Verfahrens in DIAsDEM werden Gruppierungen von identischen oder voneinander abhängigen semistrukturierten Datenbeständen, Strukturmuster und Regeln über entdeckte Beziehungen und Abhängigkeiten sein. Sie sollen dazu dienen, Anfragen an zusammengehörige oder „gleiche" Datenbestände zu richten.

Regeln zur Steuerung von Anfragen nach der Datenintegration werden in HIPED genutzt [NMO96]. Sie sind in einer deduktiven Datenbank (CORAL [RSSS94]) gespeichert und werden während der Anfrageausführung eingesetzt. Die Ergebnisregeln des Data Mining unterscheiden sich aber von solchen Regeln, indem sie nur mit einer gewissen Wahrscheinlichkeit gelten. Deshalb werden während des Integrationsprozesses und der Anfragebearbeitung Methoden des probabilistischen Schließens („probabilistic reasoning") [Pea97] benötigt. Solche Methoden werden unter anderem im Information Retrieval genutzt [RB97, RF97].

Methodik zum Einsatz von Mining-Verfahren

Der Einsatz von Mining-Verfahren ist ein komplizierter mehrphasiger Prozeß, in dem spezialisierte Kenntnisse über die Daten, die Bewertungsmaßstäbe und die Arbeitsprozesse des Anwendungsgebiets einfließen müssen. Dazu wurde die Gestaltung eines formellen Vorgehensmodells im internationalen Projekt CRISP-DM initiert [RWK+98]. CRISP-DM steht für "CRoss-Industry Standard Process for Data Mining". In seinem Rahmen werden ein generisches Referenzmodell für den Einsatz von Mining-Verfahren und ein entsprechendes Handbuch gestaltet, die anwendungs- und werkzeugabhängig sein sollen.

Wir werden die Hauptlinien von CRISP-DM selektiv als Leitfaden nutzen, entlang dem die Forschungsaktivitäten von DIAsDEM organisiert werden. Dies hat den Vorteil, daß wir Vorgehen für Aufgaben außerhalb unseren Forschungsrahmens nicht neu konzipieren müssen und notwendige Werkzeuge zielorientiert auswählen und einsetzen können.

CRISP-DM befaßt sich nicht mit dem Thema der Nutzung der Mining-Ergebnisse. In DIAsDEM wollen wir dagegen die Ergebnisse der klassischen Mining-Verfahren bearbeiten, generalisieren und zur Integration von Datensammlungen einsetzen. Unsere Methodik dazu werden wir generisch als Vorgehensmodell für die Integration von Datensammlungen mit Hilfe von Mining-Verfahren konzipieren.

Potentielle Anwendungsgebiete

In vielen Anwendungsgebieten besteht die Notwendigkeit, Daten aus heterogenen Quellen zu integrieren. Oft liegt die Strukturinformation (bzw. das Datenbankschema) nur unvollständig und gegebenenfalls auch mit Unsicherheiten behaftet vor. Typische Beispiele finden sich in der Bioinformatik. Ein umfassender Überblick über ein Chromosom oder eine Diagnose über Stoffwechselkrankheiten erfordert oft die Zusammenführung von Daten, die sich in heterogenen Datenbanken und Datensammlungen befinden. Diese Daten ergänzen sich oft oder sind sogar identisch, aber unterschiedlich repräsentiert. Die Anforderungen in diesem Anwendungsgebiet sowie die dabei auftretenden Probleme und erste Lösungsansätze sind in [HHSS98, SH98] dargestellt.

Ein weiteres Anwendungsgebiet ist das der Digitalen Bibliotheken. Für die Gestaltung von einheitlichen Anfragen und vergleichbaren Ergebnissen sind Integrationsschritte notwendig, über die unter anderem in [GCGMP97, GGM97, FS98b] berichtet wird. Die Vorteile der Entdeckung von inhaltlich relevanten Dokumenten in verschiedenen Sammlungen sind in [SS96d] anhand des Systems ARROWSMITH veranschaulicht. ARROWSMITH nutzt Mining-Verfahren, um thematisch relevante Dokumente in Sammlungen zu entdecken, *die einander nicht zitieren und unterschiedliche Fachtermini nutzen*. Da ARROWSMITH im Bereich Medizin genutzt wird, waren manche Entdeckungen (z.B. über Nebenwirkungen von Medikamenten) direkt praxisrelevant. Unser Vorhaben in DIAsDEM geht einen Schritt weiter, indem auch semistrukturierte Datenbestände außerhalb von Archiven anhand ihrer Struktur und Metadaten miteinander verglichen und verknüpft werden können.

2.2 Eigene Vorarbeiten

Vorarbeiten des Antragstellers Stefan Conrad:

Für das hier beantragte Forschungsvorhaben sind die eigenen Vorarbeiten in folgenden Bereichen relevant:

Datenbankintegration. In einem vom Land Sachsen-Anhalt geförderten Projekt (Sept.'95 – Feb.'2000; FKZ 1987A/0025 und FKZ 1987A/2527R) steht die Integration von Datenbanken in föderierten Datenbanken im Vordergrund [SCC+97, Con97]. Die fachlichen Schwerpunkte des Projektes umfassen die Entwicklung eines Schemaintegrationsansatzes (GIM) [Sch98, SS98, SC97, SS96a, SS96b], die Behandlung von Integritätsbedingungen bei der Schemaintegration [CST98, CST97] sowie die Integritätsüberwachung in föderierten Systemen [CT97, TC97].

Um dieses Projekt gruppieren sich weitere Arbeiten. So wurde ein Ansatz zur Benutzerauthentifikation in föderierten Systemen entwickelt [HS98]. Die Möglichkeiten des Einsatzes von Data-Mining–Verfahren zur Unterstützung des Schemaintegrationsprozesses wurden analysiert [HC98]. Wesentliche Bestandteile einer Erweiterung der Schemaintegrationsmethode zur Unterstützung eines inkrementellen Vorgehens wurden entwickelt [ST98]. Ferner werden Abhängigkeiten von Transaktionen in erweiterten Transaktionsmodellen modelliert und analysiert [STS98a, STS98d, STS98b, STS98c]. Eine Anwendung zentraler Bestandteile des Schemaintegrationsansatzes zum Reengineering von Klassenhierarchien ist in [SC98] beschrieben.

Modellierung von Informationssystemen. In Rahmen des BMFT-Verbundprojektes KORSO (= korrekte Software, 1991–1994; FKZ: IT 01 IS 203 D) wurde zur *Entwicklung zuverlässiger Informationssysteme* [GCD+95, CDG+93] die Spezifikationssprache TROLL *light* [HCG94, GCH93, GHC+94, CGH92] entwickelt. Mit TROLL *light* werden Informationssysteme als interagierende Objektgesellschaften modelliert. Schwerpunkt der Arbeiten waren Aspekte der Validation und Verifikation [Con96, Con95, Con94]. Anwendungen dieser Modellierungsansätze auf föderierte Datenbanken werden in [TSC97a, TSC97b] beschrieben.

Prototypische Realisierungen. Zu den zuvor genannten thematischen Schwerpunkten sowie zu weiteren Themen hat die Arbeitsgruppe Erfahrungen bei prototypischen Realisierungen gesammelt. Im Bereich der föderierten Datenbanken sind insbesondere folgende Prototypsysteme entstanden.

Die SIGMA-Bench [SEHT96] ist eine prototypische Realisierung eines Werkzeuges zur Unterstützung der Schemaintegration für föderierte Datenbanken. Die Demonstrationssoftware SIGMA-Demo [GH96, HGH97] zeigt eine einfache aus zwei heterogenen Datenbanksystemen bestehende Föderation und erlaubt es, mit dieser Föderation bei eingeschränkter Funktionalität zu arbeiten.

In Kooperation mit der Arbeitsgruppe Bioinformatik (Prof. Hofestädt, Magdeburg) ist mit der BioBench [HHSS98] im Rahmen eines studentischen Praktikums ein erster Prototyp entstanden, in dem verschiedene Datenquellen zu Stoffwechselfunktionen und -krankheiten mit einer rudimentären logischen Integration verknüpft sind. In weiteren gemeinsamen Projektarbeiten wird gegenwärtig unter anderem das Architekturkonzept der BioBench überarbeitet.

Vorarbeiten der Antragstellerin Myra Spiliopoulou:

Für das hier beantragte Forschungsvorhaben sind die eigenen Vorarbeiten in folgenden Bereichen relevant:

Datenintegration. Die Integration von Datensammlungen ist der Schwerpunkt des Forschungsvorhabens WIND, „ein Warehouse für INternet Daten“, das von der Arbeitsgruppe der HU Berlin realisiert wird [FSL97]. Im Vordergrund steht die einheitliche Gestaltung von Anfragen an strukturierten Daten und unstrukturierten, mit Metadaten versehenen Dokumenten [BSF98]. Dazu wurde eine Anfragesprache spezifiziert und Übersetzungswerkzeuge entworfen [SBEF98]. Außerdem werden Probleme der Integration auf der Middleware Ebene in Angriff genommen. Ein CORBA-basierter Lösungsansatz wird zur Zeit im Rahmen einer Diplomarbeit realisiert [SBEF98]. WIND wird demnächst mit einem Werkzeug zur Extrahierung von semistrukturierten Daten von HTML-Seiten gekoppelt [FS98b].

Dieses Vorhaben steht in engem thematischen Bezug zu weiteren Arbeiten. Die Infrastrukturunterstützung in verteilten Systemen der Schwerpunkt des Promotionsvorhabens von Herrn Arno Jacobsen [Jac98, JV98]. Herr Jacobsen ist im Rahmen des von der DFG geförderten Graduiertenkollegs 316 „Verteilte Informationssysteme“ an den Lehrstuhl angebunden. Außerdem finden Forschungsarbeiten zum Thema der Gestaltung von Informationsdiensten im WWW statt [GMS+97, GM98].

Weiterhin beteiligt sich die Arbeitsgruppe der Antragsstellerin an die Aktivitäten der Arbeitsgruppe für Umweltinformationssysteme (UIS-Gruppe) im selben Institut. WIND wurde mit Hinblick auf die Anforderungen von Umweltinformationssystemen konzipiert. Die relevanten Forschungsschwerpunkte sind das Umweltdatenmanagement in verteilten Systemen (insbesondere im Internet), der Entwurf betrieblicher Umweltinformationssysteme („BUIS“) sowie der Einsatz von Metadaten im Umweltbereich [AGM97, SFR97, RFS97, Gün98, BSF98, RG98].

Data Mining. Die Entdeckung von Navigationsmustern in Web-Traversierungsdaten ist das Thema des Forschungsvorhabens WUM, „Web Utilization Miner“. Im Schwerpunkt liegen zum einen die Entwicklung eines neuen Mining-Verfahrens, die die Eigenschaften von Weblog-Dateien berücksichtigt und zu Optimierungszwecken nutzt, zum anderen die Unterstützung des Anwendungsexperten in die Steuerung des Entdeckungsprozesses [SF98]. Bei der prototypischen Realisierung von WUM wurden unter anderem auch Probleme der Datenvorbereitung und -reinigung untersucht.

Thema des Projekts NOEMON (staatlich Griechisches Projekt, in Kooperation zwischen Uni Athen und HU Berlin) ist die Unterstützung des Experten bei der Auswahl von Mining-Verfahren. Ansätze zum Vergleich von Klassifizierungsverfahren stehen hier im Vordergrund [SKFT98]. Dafür werden Vergleichskriterien und -maßstäbe modelliert und eine Vergleichsmethodik gestaltet.

Methoden zum Vergleich von Diensten und Datenquellen anhand der Kombination von unterschiedlichen Kriterien sind in mehreren Aufgaben von DIAsDEM notwendig. Die für solche Vergleiche geeignete „Data Envelopment Analysis“ wurde für den Vergleich von Datenquellen anhand von Qualitätskriterien eingesetzt [NFS98].

Anfragebearbeitung. Im Bereich der integrierten Datenbanken und Informationssystemen müssen Probleme der Anfragegestaltung, Optimierung und koordinierter Bearbeitung in den Datenquellen gelöst werden. Außer den zuvor erwähnten Vorarbeiten über Anfragen an Informationssystemen im Rahmen von WIND [BSF98, SBEF98], befaßt sich die Antragsstellerin mit weiteren relevanten Aspekten der Anfragebearbeitung in Multidatenbanken und integrierten Informationssystemen.

Thema von [Spi96] ist die Kooperation zwischen den Optimierungsmechanismen des Multidatenbankssystems und denen der Mitgliedsdatenbanken. In [NFS98] wird das Problem der Auswahl der geeignetesten Datenquellen für die Bearbeitung einer Anfrage untersucht. Die Anfragebearbeitung von medialen Daten und deren Metadaten ist das Thema von [SHP+95]; das beschriebene System wurde im Rahmen des inzwischen abgeschlossenen ESPRIT III Projekts 6788-OSMOSE entwickelt. In [SCH95] steht das Kostenmodell für das Anfragebearbeitungssystem von [SHP+95] im Vordergrund. Kostenmodellierung ist auch das Thema von [SHV96, SF96], wobei in [SF96] der Schwerpunkt auf die Modellierung des Ressourcenverbrauchs liegt.

3 Ziele und Arbeitsprogramm

3.1 Ziele

Ziel von DIAsDEM ist die Einbindung von Altlastdaten und semistrukturierten Dokumenten in einem integrierten Informationssystem, das zur Unterstützung von Entscheidungsprozessen abgefragt werden kann.

Um Daten von unterschiedlichen Quellen in einem Informationssystem zu integrieren, muß das Schema jeder Quelle bekannt und *widerspruchsfrei* sein. Erst dann kann der Integrationsprozeß anfangen, insbesondere die Beseitigung von strukturellen und semantischen Konflikten zwischen den Quellen [Con97]. Diese Voraussetzung trifft weder bei Altlastdaten noch bei semistrukturierten Datenbeständen zu.

In Altlastdaten liegen die Integritätsbedingungen oft nicht vor und sind kaum verbindlich. In semistrukturierten Daten liegt gar kein Schema vor, und die existierenden Metadaten sind weder verbindlich, noch können zu Identifizierung von identischen Objekten genutzt werden. In DIAsDEM wollen wir diese Probleme beseitigen, um die Integration von alten und semistrukturierten Daten gemäß den klassischen Integrationsschritten zu ermöglichen. Insbesondere behandeln wir:

a. die Entdeckung von Abhängigkeiten in Altlastdaten

b. die Identifizierung von semantik-tragender Struktur in semistrukturierten Daten

c. die Nutzung der Ergebnisse von a. und b. für Datenintegration

Zur Erfüllung dieser Aufgaben werden wir Verfahren der Datenanalyse verwenden, die wir im weiteren als „Mining-Verfahren" bezeichnen. Unsere Vorgehensweise werden wir als generische Methodik konzipieren.

Entdeckung von Abhängigkeiten in Altlastdaten. Bei Altlastdaten, die einem bekannten Schema entsprechen, wollen wir funktionale Abhängigkeiten und generellere Integritätsbedingungen entdecken. Hintergrundwissen über die Semantik der schon festgelegten Metadaten und über existierende Abhängigkeiten zwischen ihnen werden wir in DIAsDEM nutzen, um Hypothesen über diese Abhängigkeiten zu gestalten und sie dann anhand der Daten zu überprüfen. Dazu kommen Klassifizierungsverfahren in Frage.

Der Anwendungsexperte bringt Hintergrundwissen über die Semantik der Daten ein. Allerdings sind die Daten häufig mit Abhängigkeiten behaftet, die Implementierungsentscheidungen entstammen und oft unintuitiv sind. Zu ihrer Entdeckung werden wir Verfahren des unüberwachtes Lernens verwenden. Diese Verfahren werden oft dazu genutzt, Beziehungen zwischen Datenwerten zu entdecken. Solche Entdeckungen wollen wir zu Regelsätzen über Abhängigkeiten generalisieren, die im Idealfall für alle Daten gelten.

Die Generalisierung von Beziehungen zwischen Werten zu Beziehungen zwischen Schemaattributen ist nicht trivial: Zum einen werden viele funktionale Abhängigkeiten von einzelnen Datenwerten verletzt. Zum anderen treten gewisse Beziehungen nur für bestimmte Werte auf und dürfen gar nicht generalisiert werden. Deshalb sollen unter den Ergebnissen der Mining-Verfahren diejenigen isoliert werden, die für sinnvollen Generalisierungen geeignet sind. Zu diesem Zweck werden wir in DIAsDEM ein Bewertungsmodell für Mining-

Ergebnisse entwerfen, in dem statistische sowie subjektive Maßstäbe berücksichtigt werden [PSM94, ST96, NS97]. Zugleich werden wir neue Maßstäbe entwerfen, die für das neue Problem der Generalisierung von Beziehungen zwischen Daten geeignet sind.

Strukturextrahierung aus semistrukturierten Daten. In DIAsDEM wollen wir jene semantische Eigenschaften von semistrukturierten Daten entdecken, durch die identische Dateninstanzen identifiziert und zusammengehörende Daten zusammengeordnet werden können. Im Rahmen von DIAsDEM konzentrieren wir uns auf semantische Eigenschaften, die durch die Anwesenheit von Strukturmerkmalen erkennbar sind. Diese Eigenschaften bezeichnen wir als „Metadaten". Es soll jedoch betont werden, daß diese Metadaten zwar strukturell den Attributen eines Objekts im objektorientierten Paradigma ähneln, es ist jedoch nicht immer möglich, in ihnen Unterstrukturen zu erkennen, die ebenfalls als semantische Eigenschaften aufweisen.

Für die Repräsentation von semistrukturierten Daten werden wir ein existierendes Graphmodell nutzen (s. unter anderem [BDHS96, AQMW97, AM98, Fau98]). Für die Auswahl zwischen den Alternativen werden wir die Ausdrucksfähigkeit jedes Modells, die Funktionalität seiner Anfragesprache und die Existenz von handhabbaren Werkzeugen zur Abbildungen der einzelnen Dateninstanzen berücksichtigen. Im Rahmen von DIAsDEM werden wir existierende, auch eigens entwickelte Infrastruktur, insbesondere Middleware zur Koppelung von Werkzeugen in einem heterogenen System [FSL97, FS98b], anwenden und gegebenfalls anpassen, um die Extrahierung der semistrukturierten Daten aus den Datensammlungen zu ermöglichen.

Auf die in das Graphmodell abgebildeten Datenobjekten werden wir Mining-Verfahren anwenden. Ziel ist es, Strukturen zu entdecken, die die Funktion eines (Sub)schemas übernommen haben, indem sie gemeinsame semantische Eigenschaften/Metadaten abbilden. Dazu kommen Verfahren zur Musterentdeckung in Frage, wie im Abschnitt „Stand der Forschung" beschrieben. Allerdings wird der Einsatz von Mining-Verfahren durch zwei Tatsachen erschwert: (i) Mining-Verfahren sind hauptsächlich für Records in Dateien oder Datenbanken konzipiert, nicht für Graphen. (ii) In vielen Fällen wird dieselbe semantische Eigenschaft durch unterschiedliche Strukturmerkmale beschrieben. Die „Labels" dieser Merkmale sind oft identisch, die Graphstruktur jedoch nicht.

Die Tatsache (i) reduziert die Anzahl der direkt einsetzbaren Verfahren zu dem Modell von [WL98], welches das OEM Graphmodell [AQMW97] voraussetzt, und dem "Web Utilization Miner" von [SF98], der den Musterentdeckungsprozeß auf einem als Baum strukturierten Datenbestand ausführt. Wir werden jedoch überprüfen, ob bis zum Beginn des Projekts weitere Verfahren verfügbar sind, die mit geringem Aufwand eingesetzt werden können. Die Auswahl des Mining-Verfahrens wird in DIAsDEM sich hauptsächlich nach der Vielfalt der unterstützten Muster richten, die zu der Qualität der erwarteten Ergebnisse erheblich beiträgt.

Die Tatsache (ii) bedeutet, daß identische Objekte wegen strukturellen Unterschieden übersehen werden können. In DIAsDEM konzentrieren wir uns auf die Erkennung identischer Objekte anhand von strukturellen *Ähnlichkeiten*. Allerdings betrachten wir angesichts dieses zweiten Problems jene Graphmodelle als besonders geeignet, die eine Abstrahierung aus den präsentationsnahen Strukturmerkmalen ermöglichen [Fau98].

Die Ergebnisse der Mining-Verfahren im Rahmen dieser Aufgabe sind Strukturmerkmale, die mit statistischen Eigenschaften über ihre Gültigkeit versehen sind. Es soll nun für jede solche Struktur entschieden werden, ob sie als Identifizierungsmerkmal fungieren kann. Diese Operation ähnelt der Entdeckung von funktionalen Abhängigkeiten zwischen Metadaten von Altlastdaten und wird entsprechend ausgeführt. Allerdings ist dafür notwendig, Verfahren zur Verifizierung von Hypothesen in klassischen Daten mit Verfahren zur Musterentdeckung zu koppeln.

Datenintegration. Die oben geschilderten Ergebnisse der Mining-Verfahren sollen für die Integration der Datenbestände in einem Informationssystem eingesetzt werden. Insbesondere werden wir die extrahierten Strukturmerkmale und funktionalen Abhängigkeiten in das Schema des Informationssystems einbauen, um (i) strukturelle und inhaltliche Konflikte auf der Schemaebene zu erkennen und zu beseitigen, und (ii) identische Dateninstanzen in den verschiedenen Datensammlungen des nun integrierten Informationssystems als solche zu erkennen und bei Anfragen zusammen zu bearbeiten.

Zur ersten Aufgabe sollen die Ergebnisse der Mining-Verfahren in einem einheitlichen Regelsatz integriert werden. Da diese Ergebnisse wahrscheinlichkeitsbedingte Regelsätze sind, brauchen wir für ihre Integration ein Repräsentationsmodell für unsicheres Wissen. Dazu werden wir in DIAsDEM Modelle und Methoden des probabilistischen Schließens [Pea97, RB97, RF97] untersuchen.

Zur zweiten Aufgabe sollen Dateninstanzen, die von verschiedenen Datensammlungen stammen und laut dem integrierten Regelsatz zusammengehören, miteinander verglichen werden. Dies gilt auch als Verifizierung des Regelsatzes, der im Rahmen der vorigen Aufgabe von den Regelsätzen der einzelnen Datensammlungen entstanden ist. Dazu werden wir Klassifizierungsverfahren einsetzen, insbesondere Entscheidungsbäume, deren Ergebnis als Regelsatz gestaltet werden kann. Diese Ergebnisregeln sind eine Verfeinerung des ursprünglichen integrierten Regelsatzes, die an den gesamten Datenbestand des integrierten Informationssystems angepaßt worden sind. In DIAsDEM werden wir diesen neuen Regelsatz für die automatische Erkennung von identischen Dateninstanzen verwenden.

Der Regelsatz am Ende der Integrationsphase ist, wie alle Ergebnisse der Mining-Verfahren, mit Wahrscheinlichkeiten versehen. In DIAsDEM streben wir danach, die Befragung des integrierten Informationssystems mit klassischen Werkzeugen, wie die Anfragesprache SQL, zu unterstützen. Deshalb werden wir einen Mechanismus entwerfen, der anhand von statistischen und anwendungsspezifischen Maßstäben den integrierten Regelsatz zu boolschen Entscheidungen nutzt.

Anwendungsszenario. Die Verfahren und Lösungen von DIAsDEM für die oben geschilderten Ziele und Aufgaben von DIAsDEM sollen anhand eines Anwendungsszenarios erprobt werden. Insbesondere ist ein repräsentativer Anwendungsbereich zu wählen, der die Integration von Altlastdaten und semistrukturierten Daten erfordert. Von diesem Bereich sind nicht nur Datenbestände zum Testen notwendig, sondern auch der Bereichsexperte, der die Semantik der Daten kennt, die inhaltliche Bedeutung und Relevanz der Mining-Ergebnisse erkennt und somit zu der Gestaltung der Bewertungsmodelle beitragen kann.

Hierzu kommt der Bereich der Bio-Datenbanken (Informations- bzw. Datensammlungen zu Enzymen, Genomen, Stoffwechselkrankheiten, etc) in Frage. In diesem Bereich sind Integrationsansätze dringend notwendig. Mit der Arbeitsgruppe Bio-Informatik von Prof. R. Hofestädt (Uni Magdeburg) bestehen bereits entsprechende fachliche Kooperationen.

3.2 Arbeitsprogramm

Das im folgenden dargestellte Arbeitsprogramm gliedert sich in fünf Arbeitsabschnitte. Arbeitsabschnitt **A** befaßt sich mit der Repräsentation von Regeln, Meta-Daten und Daten. Im Abschnitt **B** werden wir unser Anwendungsszenario definieren. **C** umfaßt die Aufgaben der Datenanalyse. In **D** werden die Ergebnisse von **C** analysiert, um die endgültigen Regelsätze für die Datensammlungen zu gestalten und die Integration durchzuführen. Der letzte Abschnitt **E** befaßt sich mit der Anfrageunterstützung im integrierten Informationssystem.

Für jeden Schritt innerhalb eines Arbeitsabschnitts ist die Arbeitsgruppe angegeben, welche die Hauptrolle zur Durchführung der Aufgaben übernimmt. Die Universität Magdeburg ist durch das Kürzel *MD* gekennzeichnet, die Humboldt-Universität zu Berlin durch das Kürzel *HU*. Aufgrund der notwendigen Kooperation, die das vorhandene Fachwissen der Antragssteller zusammenbringen soll, wird aber jede Arbeitsgruppe sich an den meisten Aufgaben beteiligen. Für manche Aufgaben ist die Beteiligung beider Arbeitsgruppen in gleichem Maß notwendig. Für diese Aufgaben ist die zuerst angegebene Arbeitsgruppe federführend.

A Ein Meta-Modell zur Repräsentation von Daten, Meta-Daten und Regeln

Hier werden wir die grundliegende Repräsentation von Daten, Meta-Daten und Regelsätzen spezifizieren, die in allen anderen Arbeitsabschnitten genutzt werden sollen.

A-1 *Repräsentation von Meta-Informationen* *MD*

Als „Meta-Informationen" bezeichnen wir kollektiv die statischen Meta-Daten zu einem Datenbestand und die Regelsätze, die zulässige Dateninstanzen beschreiben. In diesem Schritt wird ein geeignetes Modell für die Repräsentation von Meta-Informationen gestaltet. Es soll generisch genug sein, um neben einem konventionellen Schema auch semantische Eigenschaften und Strukturmerkmale von semistrukturierten Daten darzustellen. Außerdem soll es die durch Mining-Verfahren gewonnenen Regelsätze über funktionale Abhängigkeiten und Beziehungen zwischen Strukturmerkmalen unterstützen.

Wie im Abschnitt 3.1 erläutert, sind diese Regelsätze wahrscheinlichkeitsbedingt, da sie nicht für alle Daten gelten. Zur Beschreibung solcher Wahrscheinlichkeiten existieren mehrere statistische Maßstäbe, wie „support" und „confidence" [BL97], die untersucht werden sollen, um jede Regel mit einer quantitativen Abwägung ihrer *„Vertrauenswürdigkeit"* zu versehen.

Dem so gestalteten Repräsentationsmodell soll ein Regelsystem unterliegen, das probabilistisches Schließen unterstützt und in einem noch genauer zu bestimmenden Umfang auch Regelverletzungen erlaubt.

A-2 *Modellierung von semistrukturierten Daten* *HU*

Wie im Abschnitt 3.1 beschrieben, soll hier ein bereits existierendes Graphmodell zur Repräsentation von semistrukturierten Daten ausgewählt werden. Für dieses Graphmodell werden wir eine Abbildung in das Repräsentationsmodell von **A-1** entwickeln, da manche Strukturmerkmale der semistrukturierten Daten auf semantische Eigenschaften hinweisen, die zu entdecken sind.

B Das Anwendungsszenario

In diesem Arbeitsschritt werden wir mit Hilfe des Anwendungsexperten ein Szenario für DIAsDEM gestalten.

B-1 *Gestaltung des Anwendungsszenarios* *MD*

In Kooperation mit dem Anwendungsexperten werden die Datenquellen ausgewählt, die die Testdaten für DIAsDEM liefern sollen. Insbesondere werden mindestens zwei Datenquellen ausgesucht, die Altlastdaten und semistrukturierte Daten enthalten und eine partielle Überlappung ihres Inhalts aufweisen, so daß ihre Integration sinnvolle Ergebnisse für unsere Studie liefern kann. Außerdem sollen die Datenquellen repräsentativ für die Probleme im Anwendungsgebiet sein, so daß ihre Integration an sich eine nutzvolle Leistung sein wird.

B-2 *Extrahierung von Meta-Informationen und Daten* *MD*

Das Modell zur Repräsentation von Meta-Informationen und das gewählte Graphmodell werden in diesem Schritt genutzt, um die Datenbestände des Anwendungsszenarios zu beschreiben. Hier sollen Datenschemata, angenommene oder erwartete Abhängigkeiten zwischen den Daten sowie Annahmen über semantische Eigenschaften von semistrukturierten Daten registriert werden.

Um auf die Datenbestände auch zuzugreifen, werden wir existierende Mechanismen nach dem Mediator-Wrapper Paradigma auswählen und an die Umgebung von DIAsDEM anpassen. Hier kommen besonders Ansätze in Frage, in denen die Wrapper aus einer abstrakten Spezifikation der jeweiligen Datenquelle generiert werden [GPQ+97, FS98a].

In diesem Schritt werden die Daten zugleich für die Mining-Verfahren vorbereitet werden. Typischerweise werden die Daten vor der Analyse von fehlerhaften und mangelhaften Einträgen gereinigt. In DIAsDEM wollen wir jedoch gerade solche Mängel und Widersprüche durch Datenanalyse entdecken. So ist die Aufgabe hier eher, die gewählten Datenbestände an die Implementierungsspezifikationen der gewählten Mining-Werkzeuge (s. **C-1**) anzupassen.

C Datenanalyse

In diesem Arbeitsabschnitt werden die Mining-Verfahren für die Altlastdaten und semistrukturierte Daten ausgewählt, angepaßt und eingesetzt. Die Überarbeitung der Ergebnisse erfolgt im Arbeitsabschnitt **D**.

C-1 *Auswahl von Mining-Verfahren* *HU*

In diesem Schritt werden wir bestehende, frei verfügbare Mining-Verfahren auf ihre Anwendbarkeit für DIAsDEM prüfen. Insbesondere werden Verfahren zur Entdeckung von Assoziationsregeln und von Mustern untersucht, sowie Klassifizierungsmethoden für die Formulierung und Verifizierung von Hypothesen. Für die Auswahl werden wir neben den Anforderungen des spezifischen Problems „Entdeckung von Struktur und Abhängigkeiten“ auch die Ergebnisse von Vergleichsberichten [GB95] und allgemeine Richtlinien [BL97] berücksichtigen.

C-2 *Mining auf den Daten* *HU*

Dieser Schritt entspricht der klassischen Datenanalyse. Durch Einsatz von Klassifizierungsverfahren wollen wir Hypothesen über Abhängigkeiten zwischen Altlastdaten generieren und überprüfen. Hypothesen des Anwendungsexperten werden ebenfalls überprüft und gegebenenfalls verfeinert.

Durch die Entdeckung von Beziehungen zwischen Werten in den Altlastdaten wollen wir Hinweise über die Existenz weiterer Abhängigkeiten ansammeln. Dazu werden wir Verfahren zur Entdeckung von Assoziationsregeln benutzen.

Semistrukturierte Daten werden wir durch Musterentdeckungsverfahren analysiseren. Ziel ist es, strukturelle Merkmale zu entdecken, die auf semantische Eigenschaften hinweisen, und somit gewissermaßen die Rolle eines konventionellen (Sub)Schemas aufweisen.

C-3 *Bewertungsmodell für die Ergebnisse der Mining-Verfahren* *HU+MD*

Aufgabe dieses Schritts ist die Gestaltung eines Modells zur Bewertung der Ergebnisse der Datenanalyse. Insbesondere sollen die entdeckten Regelsätze nach ihrer Eignung zur Nutzung im nächsten Arbeitsabschnitt **D** beurteilt werden. Dies bedeutet, daß Regeln, die (nicht) als funktionale Abhängigkeiten generalisiert werden können oder (nicht) auf Strukturmerkmale mit semantischen Eigenschaften hinweisen, als solche erkannt werden sollen. Dazu werden wir mit Hilfe des Anwendungsexperten qualitative und quantitative Maßstäbe über die „Interessantheit“ [ST96] der Ergebnisse spezifizieren, um die Auswahl der Regelsätze für den nächsten Arbeitsabschnitt zu automatisieren.

D Analyse der Mining-Ergebnisse

In diesem Arbeitsabschnitt werden die Ergebnisse der Mining-Verfahren erarbeitet und generalisiert. Danach werden die Datenquellen anhand dieser Ergebnisse in einem einheitlichen Informationssystem integriert.

D-1 *Generalisierung von Ergebnisregeln zu Abhängigkeiten* *MD*

Aufgabe dieses Schritts ist die Generalisierung von Beziehungen zwischen Datenwerten in Abhängigkeiten zwischen Attributen von Altlastdaten. Laut dem Bewertungsmodell von **C-3** und in Interaktion mit dem Anwendungsexperten werden manche Beziehungen als Kandidaten für Generalisierung ausgewählt. Die entsprechenden Regeln werden gestaltet und ihre „Vertrauenswürdigkeit“ (s. Arbeitsabschnitt **A-1**) wird berechnet.

Die neuen Regeln werden im Regelsatz des Meta-Informationsmodells (s. Arbeitabschnitt **A-1**) integriert. Dabei ist zu prüfen, ob jede neue Regel den bestehenden Regelsatz konsistent erweitert oder ob im Rahmen des probabilistischen Schließens ein Widerspruch zustande kommt. Im zweiten Fall sollen die betroffenen Regeln und Fakten (also Dateninstanzen) identifiziert werden, um festzustellen, welche Regeln für welche Dateninstanzen anwendbar sind. Außerdem soll anhand dieser Daten die Vertrauenswürdigkeit jeder betroffenen Regel neuberechnet werden.

D-2 *Entdeckung von Abhängigkeiten zwischen Mustern* *HU*

In diesem Schritt werden häufig vorkommende Strukturmerkmale, die während der Musterentdeckung identifiziert wurden, mit semantischen Eigenschaften verknüpft. Diese Merkmale sind im allgemeinen Subgraphen mit oder ohne „Labels“ an ihren Knoten und Kanten. Nur der Anwendungsexperte kann entscheiden, ob ein solches Merkmal kennzeichnend für eine Dateninstanz oder eine semantische Eigenschaft ist. So ist diese Verknüpfung in enger Kooperation mit dem Anwendungsexperten durchzuführen.

Wie im Fall der generalisierten Regeln über Altlastdaten, sollen die mit Semantik versehenen Strukturmerkmale im Meta-Informationsmodell integriert werden. Ihre Vertrauenswürdigkeit wird anhand der Dateninstanzen berechnet, die das Strukturmerkmal aufweisen. Wie im Schritt **D-1** sollen die zu Widersprüchen führenden Regeln und Dateninstanzen identifiziert werden, und die Vertrauenswürdigkeit der betroffenen Regeln (s. **A-1**) soll neuberechnet werden.

D-3 *Integration der Regelsätze zu einem Regelsatz für das Informationssystem* *MD+HU*

In diesem Schritt sollen die Regelsätze der unterschiedlichen Datenquellen zu einem Regelsatz für das gesamte Informationssystem integriert werden. Dazu müssen zunächst Korrespondenzen zwischen den entsprechenden Meta-Informationsmodellen gefunden werden, anhand derer integrierte Regeln mit neu berechneter Vertrauenswürdigkeit entstehen können.

Obwohl in den einzelnen Regelsätzen Widersprüche schon behandelt worden sind, ist der integrierte Regelsatz im Allgemeinen nicht widerspruchsfrei. Außerdem ist es wichtig, identische oder zusammengehörende Dateninstanzen in den unterschiedlichen Datenquellen zu identifizieren. So werden wir in diesem Schritt den integrierten Regelsatz gegen das gesamte Informationssystem verifizieren, bzw. die Vertrauenswürdigkeit der Regeln überprüfen und gegebenenfalls neu berechnen, sowie neue Regeln entdecken. Dazu werden wir Klassifizierungsverfahren einsetzen.

E Anfrageunterstützung

In diesem Arbeitsabschnitt soll das in den vorigen Arbeitsabschnitten gestaltete integrierte Informationssystem um Mechanismen zur durchgängigen Anfrageunterstützung erweitert werden. Bei der Bearbeitung von Anfragen soll dabei die Vagheit und Vertrauenswürdigkeit der lokalen Daten entsprechend der im Meta-Informationsmodell bestimmten Integrationsregeln berücksichtigt werden.

E-1 *Entwicklung einer durchgängigen Anfragebearbeitung* *HU*

Ausgehend von bekannten Ansätzen zur Anfragebearbeitung in föderierten Systemen (wie etwa Multidatenbanksprachen) soll ein erweiterbares Konzept zur Anfragebearbeitung unter Berücksichtigung semistrukturierter Datenquellen entwickelt werden. Hierbei sollen Ansätze der „konventionellen" Anfragebearbeitung mit Ansätzen der Anfragebearbeitung für semistrukturierte Daten verknüpft werden.

Wir wollen in diesem Schritt kein neues Anfragebearbeitungskonzept entwickeln, sondern die bestehenden Ansätze derart erweitern, daß sie die Regeln über Abhängigkeiten, semantische Eigenschaften der semistrukturierten Daten und identische Dateninstanzen berücksichtigen. Zuerst werden wir eine Erweiterung erzielen, die die Koppelung des Anfragebearbeitungsmechanismus mit boolschen Regeln ermöglicht. Die Koppelung mit einem Regelsystem des probabilistischen Schließens wird im nächsten Schritt untersucht.

E-2 *Anfragen und Vertrauenswürdigkeit der Regelsätze* *MD*

Existierende Anfragebearbeitungssysteme für Multidatenbanken und für Sammlungen von semistrukturierten Daten basieren implizit auf einem boolesche Logiksystem. Das Meta-Informationsmodell des integrierten Informationssystems in DIAsDEM basiert aber auf probabilistisches Schließen.

Hierzu werden wir zum einem Methoden untersuchen, die Ergebnisse des probabilistischen Schließens in geeigneter Weise in boolschen Entscheidungen zu transformieren, indem wir z.B. untere Grenzen für die Vertrauenswürdigkeit anwendbarer Regeln setzen. Zum anderen werden wir aber Möglichkeiten der Erweiterung des Anfragebearbeitungssystems untersuchen, die die direkte Anwendbarkeit unserer Regel ermöglichen. Dazu werden wir zunächst die verschiedenen Ansätze aus dem Bereich *Flexible Query Answering Systems* [ACL97] untersuchen.

Diese Ansätze sind in der Regel für die Anfragebearbeitung auf einer einzelnen Datenbank ausgelegt. Hierfür ist zu prüfen, wie sich dies mit den im globalen Meta-Informationsmodell enthaltenen Integrationsregeln für Daten aus verschiedenen Datenquellen verbinden läßt. Wir erwarten hiervon eine Erweiterung der klassischen Anfragebearbeitung in föderierten Systeme um Konzepte zur Behandlung unvollständiger, vager und/oder impräziser Daten.

E-3 *Prototypische Umsetzung und Validierung* *MD+HU*

Auf der Basis der in den vorigen Arbeitsabschnitten ausgewählten Werkzeuge zur Repräsentation, Extrahierung und Anfragebearbeitung von semistrukturierten Daten, und dem integrierten Meta-Informationsmodell für unser Anwendungsszenario, werden wir abschließend eine prototypische Umsetzung der zuvor entwickelten Konzepte zur Anfragebearbeitung durchführen.

Mit diesem Prototyp sollen die entwickelten Konzepte zur Anfrageunterstützung basierend auf den Meta-Informationsmodellen des Anwendungsszenarios (Arbeitsabschnitt **B**) validiert werden.

Es folgt der graphische Zeitplan mit der Aufteilung der Arbeitsabschnitte in Schritten/Aufgaben in Abbildung 1.

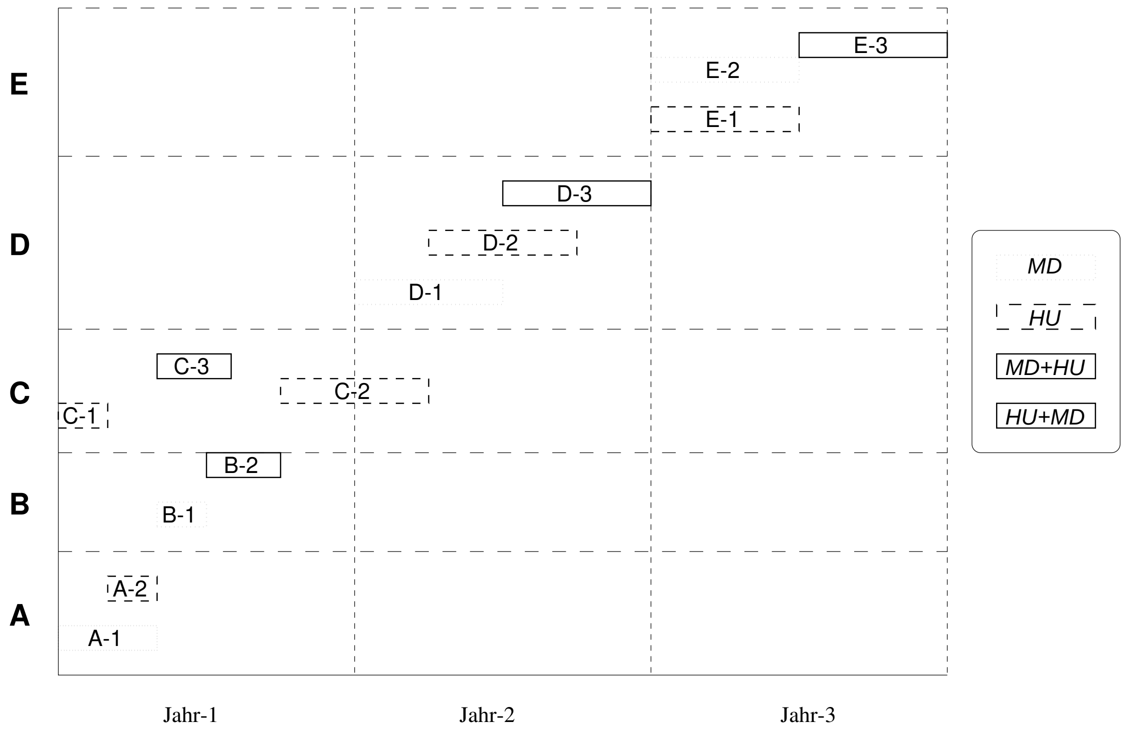

Abbildung 1: Der Zeitplan für DIAsDEM

Ausblick

Im Rahmen des hier beschriebenen Projektes können nur einige, klar umrissene Probleme behandelt werden. Für eine eventuelle Fortführung im Anschluß an dieses Projekt bieten sich aus heutiger Sicht verschiedene Schwerpunkte an. Hier wollen wir kurz nur auf einige Aspekte eingehen.

Zum einen könnten die durch Mining-Verfahren entdeckten Regeln mittels Fuzzy-Logik [ZC86, GKCR97] modelliert werden. Obwohl Fuzzy-Logik ursprünglich für die Modellierung von unscharfem Wissen konzipiert worden ist, gibt es Ansätze zu ihrer Einsetzbarkeit für Schließen bei unsicherem Wissen [DP88, HTS94]. Dazu sollten nicht unbedingt die Mining-Werkzeuge angepaßt werden, wohl aber das Regelsystem, das ihre Ergebnisse integriert. Die Behandlung von "fuzzy"-Regeln im Rahmen des Meta-Informationsmodells von DIAsDEM wirft neue Probleme auf.

Zum anderen könnte man das Projekt auch in Richtung der Inbetriebnahme eines derart aufgebauten integrierten Informationssystems fortsetzen. Es ist naheliegend, daß während des Betriebs eines solchen Informationssystems neue Erkenntnisse über die lokale Strukturierung hinzukommen könnten oder neue Korrespondenzen erkannt werden. Diese neuen Erkenntnisse müßten in die bestehende Integration einfließen, so daß hier Evolutionsprobleme zu erwarten sind. Für die Lösung solcher Evolutionsprobleme könnten erneut Mining-Verfahren eingesetzt werden. Dazu sollen Methoden untersucht werden, die die Automatisierung der Schritte zur Überarbeitung der Mining-Ergebnisse und deren Integration im Regelsatz des Meta-Informationsmodells ermöglichen.

Literatur

[Abi97] Serge Abiteboul. Querying semi-structured data. In *ICDT'97*, Jan. 1997.

[ACL97] T. Andreasen, H. Christiansen, and H.L. Larsen, editors. *Flexible Query Answering Systems*. Kluwer Academic Publishers, 1997.

[AFK97] Amihood Amir, Ronen Feldman, and Reuven Kashi. A new and versatile method for association generation. *Information Systems*, 22:333–347, 1997.

[AGM97] Hans-Knud Arndt, Oliver Günther, and Thomas Matscheroth. Metainformationssysteme als Grundlage eines BUIS-Rahmensystems. In *11th Symposium on Computer Science for Environmental Protection*. Metropolis, Marburg, 1997.

[AM98] G. Arocena and A. Mendelzon. WebOQL: Restructuring documents, databases and webs. In *ICDE'98*, 1998.

[AMM97] Paolo Atzeni, Giansalvatore Mecca, and Paolo Merialdo. To weave the web. In *VLDB'97*, pages 206–215, Athens, Greece, Aug. 1997.

[AQMW97] Serge Abiteboul, Dallan Quass, J. McHugh, and Jennifer Widom. The Lorel query language. *Digital Libraries*, 1997.

[AS95] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *ICDE*, Taipei, Taiwan, Mar. 1995.

[ATS93] Rakesh Agrawal, Imielinski T., and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD'93*, pages 207–216, Washington D.C., USA, May 1993.

[BDHS96] Peter Buneman, Susan Davidson, Gerd Hillebrand, and Dan Suciu. A query language and optimization techniques for unstructured data. In *SIGMOD'96*, pages 505–516, June 1996.

[BL97] Michael J.A. Berry and Gordon Linoff. *Data Mining Techniques: For Marketing, Sales and Customer Support*. John Wiley & Sons, Inc., 1997.

[BLN86] C. Batini, M. Lenzerini, and S.B. Navathe. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*, 18(4):323–364, 1986.

[BSF98] Steffan Baron, Myra Spiliopoulou, and Lukas Faulstich. Eine integrierte Anfrageumgebung für umweltbezogene Objekte und Dokumente. In *12th Int. Symposium "Umweltinformatik'98"*, Bremen, Germany, Sept. 1998. to appear.

[CDAR97] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In *VLDB'97*, pages 446–455, Athens, Greece, Aug. 1997.

[CDG+93] S. Conrad, G. Denker, M. Gogolla, R. Herzig, N. Vlachantonis, and H.-D. Ehrich. Zur Entwicklung zuverlässiger Informationssysteme in KorSo. In H. Reichel, editor, *Informatik — Wirtschaft — Gesellschaft, Proc. of the 23. GI-Jahrestagung (GI'93)*, Informatik aktuell, pages 464–469. Springer-Verlag, 1993.

[CGH92] S. Conrad, M. Gogolla, and R. Herzig. TROLL *light*: A Core Language for Specifying Objects. Informatik-Bericht 92–2, Technische Universität Braunschweig, 1992.

[Che96] A.L.P. Chen. Identifying Object Isomerism in Multidatabase Systems. *Distributed and Parallel Databases*, 4(2):143–168, 1996.

[Con94] S. Conrad. *Ein Basiskalkül für die Verifikation von Eigenschaften synchron interagierender Objekte*, volume 295 of *Fortschritt-Berichte (Reihe 10)*. VDI Verlag, Düsseldorf, 1994.

[Con95] S. Conrad. Compositional Object Specification and Verification. In I. Rozman and M. Pivka, editors, *Proc. of the Int. Conf. on Software Quality (ICSQ'95), Maribor, Slovenia*, pages 55–64, 1995.

[Con96] S. Conrad. A Basic Calculus for Verifying Properties of Interacting Objects. *Data & Knowledge Engineering*, 18(2):119–146, March 1996.

[Con97] S. Conrad. *Föderierte Datenbanksysteme: Konzepte der Datenintegration*. Springer-Verlag, Berlin/Heidelberg, 1997.

[CST97] S. Conrad, I. Schmitt, and C. Türker. Behandlung von Integritätsbedingungen bei Schemarestrukturierung und Schemaintegration. In K.R. Dittrich and A. Geppert, editors, *Datenbanksysteme in Büro, Technik und Wissenschaft (BTW'97)*, Informatik aktuell, pages 352–369. Springer-Verlag, Berlin, 1997.

[CST98] S. Conrad, I. Schmitt, and C. Türker. Considering Integrity Constraints During Federated Database Design. In S. M. Embury, N. J. Fiddian, A. W. Gray, and A. C. Jones, editors, *Advances in Databases, 16th British National Conf. on Databases, BNCOD 16, Cardiff, Wales, July 1998*, volume 1405 of *Lecture Notes in Computer Science*, pages 119–133, Berlin, 1998. Springer-Verlag.

[CT97] S. Conrad and C. Türker. Unternehmensweite Datenkonsistenz durch Integration bestehender Informationssysteme. In H. Krallmann, editor, *Wirtschaftsinformatik'97: Internationale Geschäftstätigkeit auf der Basis flexibler Organisationsstrukturen und leistungsfähiger Informationssysteme*, pages 345–356. Physica-Verlag, Heidelberg, 1997.

[DCH95] Surnjani Djoko, Diane J. Cook, and Lawrence B. Holder. Analyzing the benefits of domain knowledge in substructure discovery. In *KDD'95*, pages 75–80, 1995. pages 78 and 79 are switched.

[DH84] U. Dayal and H.-Y. Hwang. View Definition and Generalization for Database Integration in a Multidatabase System. *IEEE Transactions on Software Engineering*, 10(6):628–644, 1984.

[DP88] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[Dže96] Sašo Džeroski. Inductive logic programming and knowledge discovery in databases. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 117–152. AAAI Press/The MIT Press, 1996.

[Fau98] Lukas C. Faulstich. Using graph transformation techniques for integrating information from the WWW. In *Theory and Application of Graph Transformations (TAGT'98)*, 1998. to appear.

[FFLS97] Maria Fernandez, Daniela Florescu, Alon Levy, and Dan Suciu. A query language and processor for a web-site management system. In *SIGMOD Workshop on Management of Semistructured Data*, 1997.

[FKZ97] Ronen Feldman, Willi Klösgen, and Amir Zilberstein. Visualization techniques to explore data mining results for document collections. In *KDD'97*, pages 16–23, Newport Beach, CA, Aug. 1997. AAAI Press.

[FS98a] Lukas C. Faulstich and Myra Spiliopoulou. Building HyperNavigation Wrappers for Publisher Web-Sites. In *Second European Conference on Digital Libraries*, 1998. *to appear*.

[FS98b] Lukas C. Faulstich and Myra Spiliopoulou. Building HyperView wrappers for publisher web-sites. In *2nd European Conf. on Digital Libraries (ECDL'98)*, Heraclion, Greece, Sept. 1998.

[FSL97] Lukas Faulstich, C., Myra Spiliopoulou, and Volker Linnemann. WIND: A Warehouse for INternet Data. In *BNCOD'97*, London, UK, July 1997.

[GB95] J. Gama and P. Brazdil. Characterization of classification algorithms. In Pinto C. Fereira and N. Mamede, editors, *Progress in AI, 7th Portugese Conf. in AI (EPIA'95)*, pages 83–102. Springer Verlag, 1995.

[GCD+95] M. Gogolla, S. Conrad, G. Denker, R. Herzig, and N. Vlachantonis. A Development Environment for an Object Specification Language. *IEEE Transactions on Knowledge and Data Engineering*, 7(3):505–508, 1995.

[GCGMP97] Luis Gravano, Chen-Chuang K. Chang, Héctor García-Molina, and Andreas Paepcke. STARTS: Stanford proposal for internet meta-searching. In *SIGMOD'97 (experience paper)*, pages 207–218, Tucson, Arizona, June 1997.

[GCH93] M. Gogolla, S. Conrad, and R. Herzig. Sketching Concepts and Computational Model of TROLL *light*. In A. Miola, editor, *Design and Implementation of Symbolic Computation Systems, Proc. of the 3rd Int. Conf., DISCO'93, Gmünden, Austria, September 1993*, volume 722 of *Lecture Notes in Computer Science*, pages 17–32, Berlin, 1993. Springer-Verlag.

[GGM97] Luis Gravano and Héctor García-Molina. Merging ranks from heterogeneous internet sources. In *VLDB'97*, Athens, Greece, Aug. 1997.

[GH96] G. Grohmann and M. Höding. SIGMA$_{Demo}$ - Eine FDBS-Demonstrationssoftware. In W. Hasselbring, editor, *Kurzfassungen zum 2. Workshops "Föderierte Datenbanken", Dortmund, 12.-13. Dezember 1996*, number 90 in SWT Memo, pages 27–34. Fachbereich Informatik, Universität Dortmund, 1996.

[GHC+94] M. Gogolla, R. Herzig, S. Conrad, G. Denker, and N. Vlachantonis. Integrating the ER Approach in an OO Environment. In R. Elmasri, V. Kouramajian, and B. Thalheim, editors, *Proc. of the 12th Int. Conf. on the Entity-Relationship Approach (ER'93)*, volume 823 of *Lecture Notes in Computer Science*, pages 376–389, Berlin, 1994. Springer-Verlag.

[GKCR97] R. Groenemans, E.E. Kerre, D. De Cooman, and E. Van Ranst. Fuzzy Database Model Based on Quasi-Order Relations. *Journal of Intelligent Information Systems*, 8(3):227–243, 1997.

[GM98] Oliver Günther and Rudolf Müller. From GISystems to GIServices: Spatial computing in the internet marketplace. In M. Egenhofer and M. Goodchild, editors, *Interoperability in Geographic Information Systems*. Taylor & Francis, 1998. to appear.

[GMS+97] Oliver Günther, Rudolf Müller, Peter Schmidt, Bhargava H. K., and R. Krishnan. MMM: A Web-based system for sharing statistical computing module s. *IEEE Internet Computing*, 3:59–68, 1997.

[GPQ+97] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, V. Vassalos, and J. Widom. The TSIMMIS Approach to Mediation: Data Models and Languages. *Journal of Intelligent Information Systems*, 8(2):117–132, 1997.

[Gün98] Oliver Günther. *Environmental Information Systems*. Springer, Berlin/Heidelberg/New York, 1998.

[GW97] Roy Goldman and Jennifer Widom. DataGuides: Enabling query formulation and optimization in semistructured databases. In *VLDB'97*, pages 436–445, Athens, Greece, Aug. 1997.

[HC98] M. Höding and S. Conrad. Data-Mining Tasks in Federated Database Systems Design. In T. Özsu, A. Dogac, and Ö. Ulusoy, editors, *Issues and Applications of Database Technology (IADT'98), Proc. of the 3rd World Conf. on Integrated Design and Process Technology, July 6-9, 1998, Berlin, Germany*, volume 2, pages 384–391. Society for Design and Process Science, 1998.

[HCG94] R. Herzig, S. Conrad, and M. Gogolla. Compositional Description of Object Communities with TROLL *light*. In C. Chrisment, editor, *Proc. of the Basque Int. Workshop on Information Technology (BIWIT'94)*, pages 183–194, Toulouse, 1994. Cépaduès-Éditions.

[HF96] Jiawei Han and Yongjian Fu. Exploration of the power of attribute-oriented induction in data mining. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 399–421. AAAI Press/The MIT Press, 1996.

[HGH97] M. Höding, G. Grohmann, and E. Hildebrandt. Die FDBS-Demonstrationssoftware **SIGMA**$_{Demo}$. Preprint 3, Fakultät für Informatik, Universität Magdeburg, 1997.

[HHSS98] M. Höding, R. Hofestädt, G. Saake, and U. Scholz. Schema Derivation for WWW Information Sources and their Integration with Databases in Bioinformatics. In W. Litwin, T. Morzy, and G. Vossen, editors, *Second East-European Symposium on Advances in Databases and Information Systems (ADBIS'98), September 8-11, 1998, Poznan, Poland*, Lecture Notes in Computer Science, Berlin, 1998. Springer-Verlag. *To appear*.

[HM85] D. Heimbigner and D. McLeod. A Federated Architecture for Information Management. *ACM Transactions on Office Information Systems*, 3(3):253–278, 1985.

[HMS94] J. Hammer, D. McLeod, and A. Si. An Intelligent System for Identifying and Integrating Non-Local Objects in Federated Database Systems. In *Proceedings of the 27th Hawaii International Conference on System Sciences*, pages 389–407. IEEE Computer Society Press, 1994.

[HS98] E. Hildebrandt and G. Saake. User Authentication in Multidatabase Systems. In *9th International Workshop on Database and Expert Systems Applications (DEXA'98), Wien, Austria, August 1998*. IEEE Computer Society Press, 1998. *To appear*.

[HTS94] John Hale, J. Threet, and Sujeet Shenoi. A practical formalism for imprecise inference control. In *8th IFIP WG11.3 Workshop on Database Security*, 1994.

[Jac98] H.-A. Jacobsen. A generic architecture for hybrid intelligent systems. In *World Congress on Intelligent Information Systems and IEEE Fuzzy Systems*, Anchorage, Alaska, USA, May 1998. IEEE. (Invited Paper).

[JV98] H.-Arno Jacobsen and A. Voisard. Interoperable geographic information systems. In *Euro–Parallel and Distributed Systems (Euro–PDS)*, Vienna, Austria, July 1998.

[Kim95] W. Kim, editor. *Modern Database Systems: The Object Model, Interoperability, and Beyond*. ACM Press/Addison-Wesley Publishing, 1995.

[LHKK96] Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In *KDD'96*, pages 238–243, 1996.

[LMR90] W. Litwin, L. Mark, and N. Roussopoulos. Interoperability of Multiple Autonomous Databases. *ACM Computing Surveys*, 22(3):267–293, 1990.

[LNE89] James Larson, Shamkant Navathe, and Ramez Elmasri. A theory of attribute equivalence in databases with application to schema integration. *IEEE Trans. on Software Engineering*, 15:449–463, 1989.

[MT96] Heikki Mannila and Hannu Toivonen. Discovering generalized episodes using minimal occurences. In *KDD'96*, pages 146–151, 1996.

[NFS98] Felix Naumann, Johann Christoph Freytag, and Myra Spiliopoulou. Quality-driven source selection using data envelopment analysis. In *The 1998 Conference on Information Quality*, MIT, MA, Oct. 1998. to appear.

[NMO96] Shamkant Navathe, Sameer Mahajan, and Edward Omiecinski. Rule based database integration in HIPED: Heterogeneous intelligent processing in engineering design. In *Int. Symposium on Cooperative Database Systems for Advanced Applications*, Kyoto, Japan, 1996. World Scientific Press.

[NS97] Gholamreza Nakhaeizadeh and Alexander Schnabl. Development of multi-criteria metrics for the evaluation of data mining algorithms. In *KDD'97*, pages 37–42, Newport Beach, CA, Aug. 1997. AAAI Press.

[PBE95] E. Pitoura, O. Bukhres, and A. Elmagarmid. Object-Orientation in Multidatabase Systems. *ACM Computing Surveys*, 27(2):141–195, 1995.

[PE98] Mike Perkowitz and Oren Etzioni. Adaptive web pages: Automatically synthesizing web pages. In *submitted to AAAI'98*, 1998.

[Pea97] J. Pearl. Graphical Models for Probabilistic and Causal Reasoning. In A. B. Tucker, editor, *The Computer Science and Engineering Handbook.*, pages 697–714. CRC Press, 1997.

[PSM94] Gregory Piateski-Shapiro and Christopher J. Matheus. The interestingness of deviations. In *AAAI'94 Workshop Knowledge Discocery in Databases*, pages 25–36. AAAI Press, 1994.

[QRS+95] Dallan Quass, Anand Rajaraman, Yehoshua Sagiv, Jeffrey Ullman, and Jennifer Widom. Querying semistructured heterogeneous information. In *DOOD'95*, 1995. http://www-db.stanford.edu/pub/papers/querying.ps.

[RB97] T. Rölleke and M. Blömer. Probabilistic Logical Information Retrieval for Content, Hypertext, and Database Querying. In N. Fuhr, G. Dittrich, and K. Tochtermann, editors, *Hypertext - Information Retrieval - Multimedia'97: Theorien, Modelle und Implementierungen integrierter elektronischer Informationssysteme, HIM'97, Dortmund, 29.9.-2.10.1997, Proceedings*, pages 147–160. Universitätsverlag Konstanz, 1997.

[RF97] T. Rölleke and N. Fuhr. Probabilistic Reasoning for Large Scale Databases. In K. R. Dittrich and A. Geppert, editors, *Datenbanksysteme in Büro, Technik und Wissenschaft (BTW), GI-Fachtagung, Ulm, 5.-7. März 1997, Proceedings*, Informatik Aktuell, pages 118–132. Springer-Verlag, 1997.

[RFS97] Jan Röttgers, Lukas Faulstich, and Myra Spiliopoulou. Ein Verweis- und Kommunikations-Service für den betrieblichen Umweltschutz. In *BUIS'97*, pages 53–65, Berlin, Germany, Mar. 1997. Metropolis-Verlag. on german.

[RG98] Jan Röttgers and Oliver Günther. Nutzeranforderungen an Umweltinformationssysteme: Die Fallstudie VKS-Umwelt. In *12th Int. Symposium "Umweltinformatik'98"*, Bremen, Germany, Sept. 1998. to appear.

[RSSS94] Raghu Ramakrishnan, Divesh Srivastava, S. Sudarshan, and Praveen Seshadri. The CORAL deductive system. *The VLDB Journal*, 3:161–210, 1994.

[Rus98] Lucian Russell. Deductive data mining: Uncertainty measures for banding the search space. In *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1998.

[RWK+98] Thomas Reinartz, Rüdiger Wirth, Tom Khabaza, Jens H. Hejlesen, Pete Chapman, and Randy Kerber. The current CRISP-DM process model for data mining. Technical Report 98/11, Techn. Univ. Berlin, Aug. 1998. CRISP-DM home page: `http://www.ncr.dk/CRISP`.

[SBEF98] Myra Spiliopoulou, Steffan Baron, Marco Ehlert, and Lukas C. Faulstich. WIND: Ein Warehouse für INternet Daten. Technical report, Inst. für Wirtschaftsinformatik, HU Berlin, June 1998.

[SC97] I. Schmitt and S. Conrad. Restructuring Class Hierarchies for Schema Integration. In R. Topor and K. Tanaka, editors, *Database Systems For Advanced Applications'97 (DASFAA'97)*, pages 411–420. World Scientific Publishing, Singapore, 1997.

[SC98] I. Schmitt and S. Conrad. Re-engineering Object-Oriented Database Schemata by Concept Analysis. In T. Polle and K.-D. Schewe, editors, *Proc. 7th International Workshop on Foundations of Models and Languages for Data and Objects (FoMLaDO'98, Timmel, Oct. 1998)*, Informatik-Bericht. Universität Dortmund, 1998. *To appear (also expected to be published by Kluwer Academic Publ., 1999)*.

[SCC+97] G. Saake, A. Christiansen, S. Conrad, M. Höding, I. Schmitt, and C Türker. Föderierung heterogener Datenbanksysteme und lokaler Datenhaltungskomponenten zur systemübergreifenden Integritätssicherung: Kurzvorstellung des Projekts SIGMA$_{FDB}$. In K.R. Dittrich and A. Geppert, editors, *Datenbanksysteme in Büro, Technik und Wissenschaft (BTW'97)*, Informatik aktuell, pages 322–331. Springer-Verlag, Berlin, 1997.

[SCH95] Myra Spiliopoulou, Yannis Cotronis, and Michael Hatzopoulos. Query processing for multimedia applications on optical media. *Information Processing Letters*, 53:301–306, 1995.

[Sch98] I. Schmitt. *Schemaintegration für den Entwurf Föderierter Datenbanken*, volume 43 of *Dissertationen zu Datenbanken und Informationssystemen*. infix-Verlag, Sankt Augustin, 1998.

[SEHT96] I. Schmitt, A. Ebert, M. Höding, and C. Türker. **SIGMA**$_{Bench}$ – Ein Werkzeug zum Entwurf föderierter Datenbanken. In W. Hasselbring, editor, *Kurzfassungen zum 2. Workshops "Föderierte Datenbanken", Dortmund, 12.-13. Dezember 1996*, number 90 in SWT Memo, pages 19–26. Fachbereich Informatik, Universität Dortmund, 1996.

[SF96] Myra Spiliopoulou and Johann Christoph Freytag. Modelling resource utilization in pipelined query execution. In *Euro-Par'96*, pages 872–880, Lyon, France, Aug. 1996.

[SF98] Myra Spiliopoulou and Lukas Faulstich, C. WUM: A Tool for Web Utilization Analysis. In *EDBT Workshop WebDB'98*, Valencia, Spain, Mar. 1998. Springer Verlag. extended version to appear in LNCS.

[SFR97] Myra Spiliopoulou, Lukas Faulstich, and Jan Röttgers. Ein Konzept für die integrierte Recherche in unabhängigen Umweltinformationssystemen. In *11th Int. Symposium "Umweltinformatik"*, pages 327–336, Strassburg, France, Sept. 1997.

[SH98] U. Scholz and M. Höding. Erste Erfahrungen bei der Integration heterogener Informationssysteme in der Bioinformatik. In S. Conrad and W. Hasselbring, editors, *Integration heterogener Softwaresysteme IHS'98: Workshop im Rahmen der GI-Jahrestagung Informatik'98, Magdeburg, 21. September 1998*, pages 96–101, 1998. *to appear*.

[SHP+95] Myra Spiliopoulou, Michael Hatzopoulos, Franz Pentaris, Lambros Ntasios, and Fivos Vilanakis. A query processor operating over low resources on optical media. In *11th ICAST*, pages 48–55, Chicago, IL, 1995.

[SHV96] Myra Spiliopoulou, Michalis Hatzopoulos, and Costas Vassilakis. A cost model for the estimation of query execution time in a parallel environment supporting pipeline. *Computers & Artificial Intelligence*, 14(1):341–368, 1996.

[SKFT98] Myra Spiliopoulou, Alexis Kalousis, Lukas C. Faulstich, and Theoharis Theoharis. NOEMON: An intelligent assistant for classifier selection. In *FGML'98 (Treffen der GI-Fachgruppe "Maschinelles Lernen"*, Aug. 1998.

[SL90] A.P. Sheth and J.A. Larson. Federated Database Systems for Managing Distributed, Heterogenous, and Autonomous Databases. *ACM Computing Surveys*, 22(3):183–236, 1990.

[SN88] M. Schrefl and E.J. Neuhold. Object Class Definition by Generalization Using Upward Inheritance. In *Proceedings 4th International Conference on Data Engineering (ICDE'88)*, pages 4–13. IEEE Computer Society Press, 1988.

[SPD92] S. Spaccapietra, C. Parent, and Y. Dupont. Model Independent Assertions for Integration of Heterogeneous Schemas. *VLDB Journal*, 1(1):81–126, 1992.

[Spi96] Myra Spiliopoulou. A calibration mechanism identifying the optimization technique of a multidatabase participant. In *PDCS'96*, Dijon, France, Sept. 1996.

[Spi98] Myra Spiliopoulou. The laborious way from data mining to web mining. *submitted for publication*, June 1998.

[SS95] I. Schmitt and G. Saake. Managing Object Identity in Federated Database Systems. In M. Papazoglou, editor, *Proc. of the 14th Int. Conf. on Object-Oriented and Entity-Relationship Modeling (OOER'95)*, number 1021 in Lecture Notes in Computer Science, pages 400–411. Springer-Verlag, 1995.

[SS96a] I. Schmitt and G. Saake. Integration of Inheritance Trees as Part of View Generation for Database Federations. In B. Thalheim, editor, *Proc. of the 15th Int. Conf. on Conceptual Modelling (ER'96)*, number 1157 in Lecture Notes in Computer Science, pages 195–210. Springer-Verlag, 1996.

[SS96b] I. Schmitt and G. Saake. Schema Integration and View Generation by Resolving Intensional and Extensional Overlappings. In K. Yetongnon and S. Hariri, editors, *Parallel and Distributed Computing Systems: Proceedings of the ISCA International Conference (PDCS'96)*, pages 751–758. International Society for Computers and Their Applications, 1996.

[SS96c] Michael Stillger and Myra Spiliopoulou. Genetic programming in database query optimization. In *Proc. of Genetic Programming Conf. '96*, pages 388–393, Stanford, CA, July 1996.

[SS96d] Don R. Swanson and Neil R. Smalheiser. Undiscovered public knowledge: a ten-year update. In *KDD'96*, 1996.

[SS98] I. Schmitt and G. Saake. Merging Inheritance Hierarchies for Database Integration. In *Proc. of the 3rd IFCIS Int. Conf. on Cooperative Information Systems, CoopIS'98, August 20–22, 1998, New York, USA*, Los Alamitos, CA, 1998. IEEE Computer Society Press.

[ST96] Avi Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowledge and Data Eng.*, 8(6):970–974, Dec. 1996.

[ST98] I. Schmitt and C. Türker. Refining Extensional Relationships and Existence Requirements for Incremental Schema Integration. In *Proc. of the 7th Conf. on Information and Knowledge Management (CIKM'98), Washington D.C., USA*. ACM Press, 1998. *To appear*.

[STS98a] K. Schwarz, C. Türker, and G. Saake. Analyzing and Formalizing Dependencies in Generalized Transaction Structures. In T. Özsu, A. Dogac, and Ö. Ulusoy, editors, *Issues and Applications of Database Technology (IADT'98), Proc. of the 3rd World Conf. on Integrated Design and Process Technology, July 6-9, 1998, Berlin, Germany*, volume 2, pages 55–62. Society for Design and Process Science, 1998.

[STS98b] K. Schwarz, C. Türker, and G. Saake. Execution Dependencies in Transaction Closures. In *Proc. of the 3rd Int. Conf. on Cooperative Information Systems (CoopIS'98), August 20–22, 1998, New York City, USA*. IEEE Computer Science Press, 1998.

[STS98c] K. Schwarz, C. Türker, and G. Saake. Extending Transaction Closures by N-ary Termination Dependencies. In W. Litwin, T. Morzy, and G. Vossen, editors, *Second East-European Symposium on Advances in Databases and Information Systems (ADBIS'98), September 8-11, 1998, Poznan, Poland*, Lecture Notes in Computer Science, Berlin, 1998. Springer-Verlag. *To appear*.

[STS98d] K. Schwarz, C. Türker, and G. Saake. Transitive Dependencies in Transaction Closures. In B. Eaglestone, B. C. Desai, and J. Shao, editors, *Proc. of the 1998 Int. Database Engineering and Applications Symposium (IDEAS'98), July 8–10, 1998, Cardiff, Wales, UK*, pages 34–43, Los Alamitos, CA, 1998. IEEE Computer Science Press.

[TC97] C. Türker and S. Conrad. Towards Maintaining Integrity of Federated Databases. In *Data Management Systems, Proc. of the 3rd Int. Workshop on Information Technology, BIWIT'97, July 2–4, 1997, Biarritz, France*, pages 93–100, Los Alamitos, CA, 1997. IEEE Computer Society Press.

[TSC97a] C. Türker, G. Saake, and S. Conrad. Modeling Database Federations in Terms of Evolving Agents. In F. Pin, Z. W. Ras, and A. Skowron, editors, *ISMIS 1997 — Poster Proceedings of the 10th Int. Symposium on Methodologies for Intelligent Systems, Charlotte, North Carolina, October 15–18, 1997*, pages 197–208. Oak Ridge National Laboratory, 1997.

[TSC97b] C. Türker, G. Saake, and S. Conrad. Requirements for Agent-based Modeling of Federated Database Systems (Extended Abstract). In A. Cesta and P.-Y. Schobbens, editors, *ModelAge 97, Proc. of the 4th ModelAge Workshop on Formal Models of Agents, Certosa di Pontignano, Italy, January 15–17, 1997*, pages 335–343. National Research Council of Italy, Institute of Psychology, 1997.

[Wan97] Ke Wang. Discovering patterns from large and dynamic sequential data. *Intelligent Information Systems*, 9:8–33, 1997.

[WL97a] Ke Wang and Huiqing Liu. Mining nested association patterns. In *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, may 1997.

[WL97b] Ke Wang and Huiqing Liu. Schema discovery for semistructured data. In *KDD'97*, pages 271–274, Newport Beach, CA, Aug. 1997. AAAI Press.

[WL98] Ke Wang and Huiqing Liu. Discovering typical structures of documents: A road map approach. In *SIGIR'98*. ACM Press, Aug. 1998.

[ZC86] A. Zvieli and P. P. Chen. Entity-Relationship Modeling and Fuzzy Databases. In *Proceedings of the Second International Conference on Data Engineering, February 5-7, 1986, Los Angeles, California, USA*, pages 320–327, Los Alamitos, 1986. IEEE Computer Society.

[ZEMK97] Oren Zamir, Oren Etzioni, Omid Madani, and Richard M. Karp. Fast and intuitive clustering of web documents. In *KDD'97*, pages 287–290, Newport Beach, CA, Aug. 1997. AAAI Press.