Datenbank- und Visualisierungstechnologie in der Informationsfusion

Gunter Saake[1] Kai-Uwe Sattler[1]
Daniel Keim[2]

[1] Fakultät für Informatik, Universität Magdeburg
Postfach 4120, 39016 Magdeburg
{saake|kus}@iti.cs.uni-magdeburg.de

[2] Institut für Informatik, Universität Halle-Wittenberg
Kurt-Mothes-Straße 1, 06099 Halle/Saale
keim@informatik.uni-halle.de

Zusammenfassung

In vielen Anwendungsbereichen besteht die Aufgabe, Daten oder Informationen aus verschiedenen, zum Teil heterogenen Quellen zu kombinieren, zu verdichten und daraus Informationen einer neuen Qualität abzuleiten. Wesentliche Kernfunktionen dieses als *Informationsfusion* bezeichneten Prozesses sind dabei durch Methoden der Datenintegration und der Datenanalyse / Data Mining bereitzustellen. Die gewachsenen Strukturen der heute genutzten Informationsquellen und die damit im Zusammenhang stehenden Probleme wie Heterogenität, Inkonsistenz oder Ungenauigkeit der Daten sind mit den aktuell verfügbaren Techniken nur bedingt beherrschbar. Ausgehend vom aktuellen Stand der Forschung diskutiert der vorliegende Beitrag Anforderungen an Datenbank- und Visualisierungstechnologien aus Sicht der Informationsfusion und zeigt aktuelle Forschungsrichtungen auf.

1 Motivation

Der heutige Stand der Datenbanktechnologie ermöglicht die effiziente Speicherung und Verwaltung von Datenbeständen mit unterschiedlichen Strukturen im Giga- und Terabyte-Bereich. Gleichzeitig erlauben moderne Kommunikationsmedien wie das Internet den Zugriff auf weltweit verteilte Informationen. Als eine Folge dieser Entwicklungen sehen sich jedoch viele Anwender einer wachsenden Informationsflut ausgesetzt, die das Auffinden von relevanten Informationen erschwert. Verschärfend wirkt dabei auch, daß viele Datenbestände zum Teil anarchisch gewachsen sind und demzufolge heterogen (sowohl bezüglich der Struktur als auch der Repräsentation) sind sowie Redundanzen und Inkonsistenzen enthalten. Die Integration externer Informationen (z.B. aus dem World Wide Web) eröffnet zwar neue Nutzungspotentiale für die unternehmensinternen Informationssysteme,

ist aber gleichzeitig mit neuen Fragestellungen verbunden, wie z.B. der Gewährleistung von Aktualität und Vertrauenswürdigkeit von Informationen, dem effiziente Zugriff auf weltweit verteilte Quellen oder der Aufbereitung von unzureichend strukturierten Daten.
Darüber hinaus enthalten die integrierten Datenbestände oft auch Informationen, die nicht explizit abgelegt sind, sondern sich in Form von Abhängigkeiten, Beziehungen oder Mustern über die einzelnen Quellen hinweg repräsentieren. Bei der Suche und Extraktion dieser impliziten oder „versteckten“ Informationen versagen jedoch klassische Anfragetechniken aus dem Datenbankbereich. Eine wichtige Anforderung an Informationssysteme ist demzufolge eine (semi-)automatisierte und intelligente Transformation der Daten in „nützliche“ Informationen. Der Begriff der Transformation umfaßt dabei verschiedene Aspekte, wie Integration, Filterung, Analyse und Aufbereitung der Daten mit dem Ziel des Aufdeckens und der Repräsentation des impliziten Wissens.
Vor diesem Hintergrund sind neben dem Datenmanagement und der Datenintegration die Bereiche Data Mining und Data Fusion Gegenstand aktueller Forschung. *Data Mining* beschäftigt sich als Kern des Prozesses der Wissensfindung in Datenbanken (Knowledge Discovery in Databases - KDD) mit der Suche nach Mustern und Abhängigkeiten in Daten. *Data Fusion* beschreibt die Kombination und Interpretation von Daten aus verschiedenen Quellen. Der Einsatz dieser Techniken im Rahmen von Informationssystemen eröffnet neue Möglichkeiten hinsichtlich der Analyse und Verdichtung großer, heterogener Datenbestände. Für diesen Prozeß der Integration und Interpretation von Daten aus verschiedenen heterogenen Quellen sowie die darauf aufbauende Konstruktion von Modellen für einen bestimmten Problembereich mit dem Ziel der Gewinnung von Informationen einer neuen, höheren Qualität wird der Begriff *Informationsfusion* verwendet.
Aus der Zielstellung der Informationsfusion resultieren Anforderungen an Methoden und Techniken aus dem Datenbankbereich sowie der Datenvisualisierung. Diese betreffen vor allem die effiziente Datenanalyse und -aufbereitung in verteilten, heterogenen Quellen und die Behandlung unzureichend strukturierter, inkonsistenter oder vager Informationen. Im vorliegenden Beitrag werden ausgehend vom aktuellen Forschungsstand wichtige Anforderungen aufgezeigt und potentielle Entwicklungsrichtungen diskutiert.

2 Stand der Forschung

Informationsfusion ist ein interdisziplinäres Gebiet, das auf Methoden und Techniken verschiedener Bereiche, wie z.B. Datenbanken, Statistik, Maschinelles Lernen, Soft Computing oder Visualisierung zurückgreift. Nachfolgend soll der Stand der aktuellen Forschung zu dieser Thematik im wesentlichen aus Datenbanksicht skizziert werden. Die betrachteten Kerngebiete sind dabei Datenintegration und -management, Data Mining und Data Fusion.
Datenintegration und -management. In der Literatur wird Daten(bank)integration oft mit Schemaintegration gleichgesetzt. Eine ganze Reihe von Ansätzen wurden für die Schemaintegration entwickelt. Verschiedene Überblicke darüber sind z.B. in [BLN86, PBE95, Con97] zu finden. Ein zentrales Anliegen der Datenbankintegration ist die Überwindung der Heterogenität auf verschiedenen Ebenen. Im Mittelpunkt dieser Arbeiten stehen in der Regel die Heterogenitäten, die durch unterschiedliche Datenmodelle und unterschiedliche

Modellierung beim Datenbankentwurf entstehen. Die übliche Vorgehensweise bei der Integration läßt sich wie folgt (idealisiert) darstellen. Zunächst werden die zu integrierenden Datenbankschemata in ein gemeinsames Datenmodell transformiert, um die Heterogenität auf Datenmodellebene zu beseitigen. Anschließend müssen die übereinstimmenden Teile der Schemata identifiziert werden, damit dann die eigentliche Integration durchgeführt werden kann. Aufgrund der häufig anzutreffenden Heterogenität in der Modellierung desselben oder eines ähnlichen Sachverhaltes muß man sich bei diesem Schritt insbesondere mit dem Vergleich unterschiedlicher Modellierungen beschäftigen. Eine weitverbreitete Schemaarchitektur, die dieses Vorgehen unterstützt, ist die 5-Ebenen-Schema-Architektur [SL90]. Neben der Betrachtung auf Schemaebene ist für den Betrieb eines föderierten Datenbanksystems auch die Realisierung von Datenzugriffsschnittstellen zu den einzelnen Systemen wichtig. Hier gibt es bereits verschiedene Prototypsysteme, in denen Datenbankadapter entwickelt wurden (z.B. in IRO-DB [GGF+96]).
Für die Anfragebearbeitung in föderierten Datenbanken (siehe auch [MY95]) müssen bereits bei der Schemaintegration Abbildungsinformationen zwischen den lokalen und dem integrierten Schema festgelegt werden. Während dadurch die Anfragebearbeitung prinzipiell möglich wird, sind der Optimierung aufgrund der Heterogenität der Systeme erhebliche Grenzen gesetzt. Insbesondere müßte eine globale Instanz möglichst vollständiges Wissen über die lokalen Anfrageoptimierungsstrategien haben oder diese direkt steuern können. Aktuelle Bemühungen zu dieser Thematik sind u.a. auch auf die Optimierung von Anfragen über autonome Internet-Datenbanken (*fusion queries*) gerichtet [YPAGM98].
Für die Beantwortung von Anfragen ist die Datenqualität bei der Integration von hoher Bedeutung. Obwohl dieses Problem schon lange bekannt ist, gibt es bisher nur wenige Arbeiten, z.B. [Ger98], dazu. Dieses Problem spielt auch im Bereich des Data Warehousing [Inm96] eine erhebliche, oft aber vernachlässigte Rolle. Ein Data Warehouse kann dabei als eine materialisierte Sicht oder Integration von operativen Datenbeständen betrachtet werden, auf der z.B. entscheidungsunterstützende Auswertungen durchgeführt werden sollen, ohne das operative Geschäft zu beeinträchtigen. Ein Data Warehouse unterscheidet sich in dieser Sichtweise von einem föderierten Datenbanksystem vorrangig dadurch, daß der integrierte Datenbestand materialisiert wird, um einen effizienten Zugriff datenintensiver (OLAP-) Anwendungen zu unterstützen. Änderungen der lokalen, operativen Daten können sich so aber nicht unmittelbar auf den integrierten Datenbestand auswirken. Für Anwendungen mit der Anforderung nach möglichst aktuellen Daten oder einem erheblich anderem Zugriffsverhalten bietet sich die logische Integration in einem föderierten Datenbanksystem an [Con97].
Data Mining/KDD. Der Begriff *Knowledge Discovery in Databases (KDD)* wird im wissenschaftlichem Umfeld als „der nichttriviale Prozeß der Identifikation gültiger, neuer, potentiell nützlicher und verständlicher Muster in Datenbeständen“ [FPSS96] definiert. *Data Mining* bezeichnet in diesem Zusammenhang den Teilschritt der Suche und Bewertung von Hypothesen. Im kommerziellen Bereich wird dagegen Data Mining häufig als Synonym für KDD verwendet. KDD ist ein iterativer und interaktiver Prozeß, der die folgenden Schritte umfaßt: (1) Festlegung von Problembereich und Zielen, (2) Datensammlung und -bereinigung, (3) Auswahl und Parametrisierung der Analysefunktionen und -methoden, (4) Data Mining, (5) Bewertung und Interpretation der Ergebnisse sowie (6) Nutzung des gefundenen Wissens.

Datenbanktechnologie wird hierbei insbesondere in den Schritten 2 und 3 eingesetzt. In Abhängigkeit von der Analyseaufgabe kommen verschiedene Methoden des Data Mining zum Einsatz. Die wichtigsten Klassen dieser Verfahren sind u.a. [FPSS96, CHY96, Wro98]:

- *Erkennung von Abhängigkeiten:* Diese Verfahren ermitteln statistische Abhängigkeiten zwischen Variablen der relevanten Datensätze. Als Ergebnis werden Assoziationsregeln [AS94] oder auch Wahrscheinlichkeitsnetze geliefert.

- *Klassifikation:* Klassifikationsverfahren zielen auf die Zuordnung von Objekten zu verschiedenen vorgegebenen Klassen ab, wobei das Klassifikationsmodell anhand einer Beispielmenge (Trainingsset) der Datenbank ermittelt wird.

- *Clustering:* Beim Clustering werden ähnliche Objekte in neu gebildete Kategorien eingeordnet, so daß die Ähnlichkeiten der Objekte innerhalb einer Kategorie möglichst groß und zwischen den Kategorien gering sind [Fis95].

- *Generalisierung:* Dies beinhaltet Methoden zur Aggregation und Verallgemeinerung großer Datenmengen auf einer höheren Abstraktionsebene. Oft werden diese Verfahren bei der interaktiven Datenexploration angewendet [HCC92].

- *Sequenzanalyse:* Diese Verfahren dienen zur Suche nach häufig auftretenden Episoden oder Ereignisfolgen in Datenbeständen, denen eine (z.B. zeitliche) Ordnung der einzelnen Datensätze zugrundeliegt [MTV95].

Neben der Verarbeitung einfacher, relationaler Daten gewinnt die Analyse von Textdokumenten (*Document/Text Mining*), Bilddatenbanken (*Image Mining*), geographischen Daten (*Spatial Data Mining*) und Informationen aus dem World Wide Web (*Web Mining*) zunehmend an Bedeutung. Für eine weitergehende Diskussion konkreter KDD-Verfahren sei an dieser Stelle auf die Literatur [SHF96, HMPU97] verwiesen.
Datenfusion. Der Begriff der Datenfusion wird gegenwärtig in verschiedenen Anwendungsgebieten in teilweise unterschiedlicher Bedeutung verwendet. Grundsätzlich wird darunter jedoch die Kombination von Daten aus verschiedenen (heterogenen) Quellen verstanden. Konkrete Problemstellungen sind u.a. [LK95, PGV95, AZ98]: *Multi-Sensor Fusion*, die sich mit der Verbindung von Meßdaten verschiedener Geräte (z.B. in industriellen oder militärischen Bereichen) befaßt, *Multiple Source Interrogation* mit dem Ziel der Zusammenführung von Informationen aus Daten- oder Wissensbanken und *Image Fusion* als Kombination von Bildern einer Szene mit dem Ziel einer verbesserten Interpretation (z.B. Röntgenaufnahmen oder Satellitenbilder).

3 Anforderungen der Informationsfusion

Aufgrund der vielfältigen Problemstellungen und Anwendungsfelder der Informationsfusion und der damit verbundenen Anforderungen werden wir uns im weiteren auf die Fusionierung von Daten bzw. Informationen aus Datenbanken beschränken. Die dafür notwendigen Funktionen eines Softwaresystems lassen sich wie folgt beschreiben:

- *Datenzugriff:* Zunächst ist der transparente Zugriff auf Daten aus unterschiedlichen Quellen zu realisieren. Dies schließt die Verwendung von Datenbank-Gateways zur Verbergung der Heterogenität ebenso ein wie die Verarbeitung von Dateien mit vorgegebener Struktur (semistrukturierte Daten), wobei der Zugriff über entsprechende Protokolle (z.B. HTTP) erfolgen kann. Weiterhin ist zu diesem Funktionsbereich die Verarbeitung und Optimierung von Anfragen zu zählen.

- *Datenintegration:* Für die Daten aus den einzelnen Quellen ist eine integrierte Sicht zu schaffen, die die Daten in einem homogenen Modell präsentiert und dabei Konflikte auf Schema- und Instanzebene behebt. Weiterhin sind quellenübergreifende Beziehungen zu repräsentieren und in geeigneter Weise zu verwalten.

- *Analyse und Verdichtung:* Durch das Extrahieren von Zusammenhängen und Abstraktionen, durch Filterung und Verdichtung der Daten sind Informationen einer neuen Qualität zu gewinnen. Die Definition der „neuen Qualität“ ist dabei abhängig von der konkreten Anwendung. Mögliche Repräsentationen für diese Informationen sind generalisierte Aggregationen und Assoziationen, Cluster und Klassen.

- *Präsentation und Weiterverarbeitung:* Die gewonnenen Informationen sind entsprechend der Problemstellung zu präsentieren bzw. zur Weiterverarbeitung bereitzustellen [KK96].

- *Repräsentation von Metainformationen:* Eine wesentliche Voraussetzung für die Fusion ist das Vorhandensein von Informationen über die Datenquellen, die zu fusionierenden Objekte und den Problembereich. Diese Metainformationen sind durch das System zu verwalten und im Verlauf des Fusionsprozesses sukzessive anzupassen bzw. zu erweitern.

Diese Funktionsbereiche sind durch eine Infrastruktur bereitzustellen, die die Basis für unterschiedliche Anwendungen der Informationsfusion bildet.
Aus der Gesamtsicht eines Systems zur Informationsfusion stellen sich darüber hinaus noch folgende Forderungen:

Unterstützung einer interaktiven und iterativen Arbeitsweise. Aufgrund der Komplexität des Fusionsprozesses und des Volumens der zu verarbeitenden Daten ist eine schrittweise, interaktive Durchführung der Fusion notwendig. So kann zunächst mit wenigen, ausgewählten Daten eines relevanten Teilschemas die Anwendbarkeit bestimmter Methoden und die zu erwartende Qualität der Ergebnisse mit geringem Aufwand abgeschätzt werden, bevor die Fusion über den gesamten integrierten Datenbestand erfolgt. Weiterhin ist eine schrittweise Fokussierung auf bestimmte Datenbereiche wünschenswert. Aus der Beobachtung der Kosten für Datenzugriff und -analyse ergibt sich außerdem die Möglichkeit einer Optimierung des Laufzeitverhaltens durch geeignete Caching-Strategien, die Verwendung speziell angepaßter Indexstrukturen, die Parallelisierung der Verarbeitung oder auch durch die Umordnung von Analyseoperationen.

Anpassungsfähigkeit und Erweiterbarkeit des Systems. Auch wenn ein Fusionssystem für einen konkreten Anwendungsfall entwickelt wird, können nicht immer die zu unterstützenden Datenquellen und die benötigten Methoden vorausbestimmt werden. Die Erweiterbarkeit des Fusionssystems um neue Methoden (z.B. durch Plugins [WWSE96]) und Datenquellen ist daher ein weiteres wichtiges Kriterium.

Intelligente Nutzerunterstützung bei Auswahl und Anwendung der Fusionsmethoden. Der Einsatz der verschiedenen Integrations-, Fusions- und Analysemethoden setzt tiefe Kenntnisse über den Problembereich, die Struktur und Semantik der Daten sowie die Methoden selbst voraus. Die Auswahl, Kombination und gegebenenfalls Parametrisierung der Methoden ist durch geeignete Techniken (z.B. Nutzung von Metainformationen und Wissen zum Problembereich, Voranalyse der Daten, Einsatzkriterien) zu unterstützen.
Mit den heute verfügbaren Techniken sind einige dieser Anforderungen bereits erfüllbar. Forschungsbedarf besteht aber noch bei der Verbindung dieser Techniken, wie z.B. der Fusion von Daten unterschiedlicher Qualität und Repräsentation, der effizienten Analyse verteilter, heterogener Quellen und der kombinierten Anwendung verschiedener Analyse- und Fusionsmethoden.

4 Anforderungen an die Datenbanktechnologie

Aus der Zielstellung der Informationsfusion und der beschriebenen Funktionalitäten leiten sich Anforderungen an Entwicklungen im Bereich Datenbanken, Visualisierung und KDD ab. Für den Datenbankbereich beinhaltet dies die folgenden Aufgabenstellungen:

- *Intelligente Unterstützung des Integrationsprozesses:* Für viele Einsatzfälle ist die Integration der Schemata der einzelnen Quellen in ein globales Schema ein komplexer Prozeß, der nicht vollständig automatisierbar ist. So sind semantische und strukturelle Konflikte der Ausgangsschemata zu beseitigen und unterschiedliche Klassenhierarchien zu integrieren. Es werden daher Werkzeuge benötigt, die diese Schritte unterstützen und dabei auch die Semantik der Daten einbeziehen. Darüber hinaus sind Aspekte der Qualität der zu integrierenden Daten zu berücksichtigen, da diese die Ergebnisse der Fusion nachhaltig beeinflussen.

- *Realisierung eines effizienten Datenzugriffs:* Zur Analyse großer Datenbestände über verschiedene Quellen hinweg sind effiziente Zugriffsmechanismen notwendig. Speziell für verteilte, heterogene Quellen sind geeignete Indexstrukturen aufzubauen und spezielle Caching- oder Replikationsstrategien zu verfolgen. Weiterhin sind die spezifischen Anforderungen der Analysemethoden hinsichtlich der Zugriffsschnittstellen (z.B. satzorientierte oder navigierende Zugriffe) zu berücksichtigen.

- *Integration semistrukturierter Daten:* Nicht zuletzt durch die Verbreitung des World Wide Web liegen viele Informationen in nur unzureichend strukturierter Form, wie z.B. in HTML-Dateien, vor. Die Integration dieser semistrukturierten Daten, die effiziente Aufbereitung und Repräsentation sowie die damit verbundene Möglichkeit der Anfragebearbeitung stellt eine weitere wichtige Aufgabe dar.

- *Gewinnung von Metainformationen:* Informationen, die die Semantik und Qualität der Daten beschreiben, bilden eine wichtige Basis für die Fusion. Sofern diese Metainformationen nicht vorliegen, müssen sie aus den Daten extrahiert oder mit Hilfe des Nutzers erfaßt werden.

Ein Teil dieser Themen sind aktueller Forschungsgegenstand z.B. im Umfeld föderierter Datenbanken. Aus den aufgeführten Aufgabenstellungen lassen sich konkrete Anforderungen an Datenbankmanagementsysteme (DBMS) ableiten:

- Das eingesetzte DBMS muß einen offenen Optimierer haben, um Fusionsmethoden zusammen mit DB-Operationen optimieren zu können. So werden in Fusionsprozessen in der Regel (statistische) Aufbereitungsschritte im Wechsel mit Filterungsschritten zur Selektion eingesetzt, die durch DBMS-Anfragen effizient unterstützt werden können. Hier muß eine übergreifende Optimierung erfolgen.
- Das DBMS muß Funktionalität zur Unterstützung der Integration externer Daten anbieten, die über einfache Import/Export-Routinen hinausgeht.
- Das DBMS sollte Funktionen zur Unterstützung des Rankings bzw. der Qualitätsbewertung von Anfrageergebnissen besitzen (also müssen in gewissem Umfang Techniken des Information Retrieval in das DBMS verlagert werden).
- Das DBMS muß eine offene Softwarearchitektur mit einem zugänglichen Repository für Metainformationen besitzen, um Fusionsmethoden einbetten zu können.
- Für die (physische) Anfragebearbeitung sollten neue Techniken, wie etwa die Indexgenerierung *on the fly*, aber auch die Nutzung und Integration spezieller vorhandener Indexstrukturen in den verschiedenen Datenquellen verfügbar sein.
- Viele statistische Methoden benutzen eine zufällige Auswahl von Datensätzen als ersten Schritt zur Initialisierung, bevor der gesamte Datenbestand analysiert wird. Ein derartiges *Sampling* zur Generierung einer zufälligen Stichprobe wird von kommerziellen DBMS in der Regel nicht unterstützt.

5 Anforderungen an die Visualisierungstechnologie

Bei der Informationsfusion sollen Datenbestände aus verschiedenen Anwendungsbereichen in Zusammenhang gebracht und ausgewertet werden. Dafür ist eine schnelle und effektive Exploration der zu fusionierenden Datenbestände notwendig. Visualisierungstechniken können dabei helfen, indem sie

- eine interaktive Datenexploration mit Hilfe einer direkten Visualisierung der Daten ermöglichen,
- automatische Data Mining Verfahren durch ein direktes Feedback unterstützen, und
- Techniken für eine effektive Interaktion mit den Daten zur Verfügung stellen.

Alle drei Aspekte sind für einen effizienten und effektiven Informationsfusionsprozeß von großer Bedeutung.
Bei einer direkten Visualisierung der Daten muß in Abhängigkeit von der Größe der Datenmenge und ihrer Dimensionalität eine geeignete Repräsentationsform gewählt werden. Beispiele für Techniken, die eine Visualisierung sehr großer multidimensionaler Daten erlauben, sind die pixel-basierten Techniken, die jeden Datenwert durch ein farbiges Pixel darstellen [Kei00]. Diese Techniken müssen für den Zweck der Informationsfusion, d.h. um die heterogenen Daten verstehen und ihr Informationsfusionspotential erkennen zu können, weiterentwickelt werden.
Aufgrund der sehr großen Datenmengen sowie der komplexen Ähnlichkeitsmaße führt eine direkte Visualisierung der Daten nicht immer direkt zu guten Clusterungsergebnissen. Um die bestmöglichen Ergebnisse zu erhalten, ist deshalb eine Kombination mit automatischen Verfahren notwendig. Ziel ist es, die Stärken visueller Verfahren wie direktes Feedback und gute Benutzerinteraktion mit den Stärken automatischer Verfahren zu kombinieren. Insbesondere hängt die Effektivität automatischer Verfahren von einer Vielzahl von Parametern ab, die nur schwer a-priori einstellbar sind und von einer Datenmenge zur nächsten signifikant variieren. Visualisierungstechniken können hier dazu beitragen, die Daten sowie die Ergebnisse der automatischen Clusterungsverfahren besser zu verstehen und den Benutzer beim Einstellen der Parameter zu unterstützen [HKW99]. Die Visualisierungstechniken erlauben es dem Benutzer, die Signifikanz der entdeckten Information (Cluster, Abhängigkeiten) abzuschätzen, den Einfluß von Parametern automatischer Verfahren zu verstehen und diese Parameter zu tunen. Visualisierungstechniken erlauben zudem ein abstrakteres Verständnis der Daten und ermöglichen, informales Hintergrundwissen in den Explorationsprozeß mit einfließen zu lassen. Durch die Interaktion mit einem Benutzter wird insgesamt die Qualität und Effektivität der Fusion entscheidend verbessert.
Zusätzlich zur direkten Visualisierung der Daten und der Integration mit automatischen Verfahren ist ferner die Verwendung von Interaktionstechniken notwendig. Die Effektivität der einzelnen Visualisierungstechniken kann beispielsweise durch eine interaktive Kombination verschiedener Visualisierungstechniken noch deutlich gesteigert werden. Wichtig ist dabei, die Verbindung zwischen den einzelnen Visualisierung auch innerhalb der Visualisierung darzustellen. Dazu können sogenannte *Linking und Brushing* Techniken [War94] verwendet werden, die es dem Benutzer erlauben, interaktiv einen Teil der Visualisierung zu markieren. Die ausgewählten Daten werden dann auch in den anderen Visualisierungen entsprechend markiert. Dadurch können zum Beispiel höher-dimensionale Strukturen mit Hilfe mehrerer zwei-dimensionaler Visualisierungen dargestellt werden.
Eine weitere Interaktionstechnik sind die multi-level Visualisierungen, die eine Visualisierung der Daten auf verschiedenen Abstraktionsstufen, einen transparenten Wechsel zwischen den Visualisierungen sowie einen transparenten Zugriff auf die Originaldaten erlauben. Die Möglichkeit eines transparenten Wechsels zwischen verschiedenen Darstellungsformen wird auch als *Detail-on-Demand* bezeichnet [Shn96]. Zu einer weiteren Effektivitätssteigerung bei der Datenexploration können ferner *Verzerrungstechniken (Distortion Techniques)* verwendet werden [LA94]. Dabei werden die Daten je nach Abstand zum aktuellen Fokus so 'verzerrt', daß die Daten im Zentrum mit hohen Detaillierungsgrad dargestellt werden, der mit steigendem Abstand vom Zentrum kontinuierlich abnimmt.

Alle genannten Aspekte - direkte Visualisierung, Visualisierungsunterstützung von automatischen Data Mining Verfahren sowie die Kombination mit Interaktionstechniken - sind eine wichtige Voraussetzung für eine effektive Informationsfusion.

6 Forschungsaktivitäten

Wichtige der in den vorangegangenen Abschnitten beschriebenen Problemstellungen werden gegenwärtig im Rahmen der DFG-Forschergruppe „Workbench für die Informationsfusion" an der Fakultät für Informatik der Universität Magdeburg bearbeitet. Die Vision des Gesamtprojektes ist die Entwicklung einer *Workbench* zur Realisierung effizienter Informationssysteme mit anwendungsspezifischen Analyse- und Fusionsmethoden. Dies beinhaltet die Bereitstellung einer Infrastruktur zur Datenhaltung und zum Zugriff auf heterogene Quellen, eine Bibliothek von Methoden zur Informationsanalyse und -fusion, sowie Werkzeuge zur Akquisition von Wissen und zur Visualisierung der Daten und Informationen. Für verschiedene Anwendungsbereiche sollen diese zunächst generischen Komponenten in geeigneter Weise angepaßt, optimiert und im Rahmen eines Informationssystems zusammengesetzt werden. Dies schließt die Unterstützung bei Auswahl und Einsatz der Fusionsmethoden, die Konfigurierung der Datenhaltungs- und Zugriffskomponenten sowie den anwendungsspezifischen Aufbau von Ontologien und Metainformationen ein. Hierfür werden Mechanismen zur Komposition und Adaption von Software-Bausteinen benötigt. Weiterhin ist die Erweiterbarkeit der Systeme hinsichtlich neuer Methoden und/oder Informationsquellen zu berücksichtigen.
Im einzelnen umfaßt die Forschergruppe folgende Projekte:

- *Föderations- und Integrationsdienste für die Informationsfusion*
  Mit diesem Forschungsvorhaben werden Techniken zum effizienten Zugriff auf heterogene Datenquellen als Basis einer generischen, in Abhängigkeit von der Anwendung konfigurierbaren Informationsfusionstechnologie entwickelt. Weiterhin soll diese Projekt den konzeptionellen und softwaretechnischen Rahmen für das Gesamtvorhaben schaffen.

- *Nachhaltige Informationsfusion: Aktives Lernen*
  In disem Projekt werden Techniken zur Aufwertung von Lernverfahren durch aktive Komponenten untersucht, mit Schwerpunkt auf den logikorientierten Lernverfahren. Dies umfaßt Techniken zur Datenauswahl aus vorgegebenen Relationen mit Hilfe von aktiven Sampling, Techniken zur Auswahl zusätzlich zu nutzender Relationen sowie Revisionstechniken zur Modifikation bestehender Modelle auf der Basis eines Abgleichs mit veränderten Daten.

- *Integration von Werkzeugen zur Wissensakquisition*
  Im Rahmen dieses Projektes soll ein Konzepts des *document mining* zur Überführung von Wissen aus (natürlichsprachlichen) Dokumenten in formale Strukturen erarbeitet werden. Hierzu werden Techniken der Integration von partiellem und redundantem Wissen in eine einheitliche Gebietsmodellierung untersucht sowie methodische und

technischen Hilfen zur Überwindung des *knowledge acquisition bottleneck* beschrieben.

- *Aspektorientierte Laufzeitumgebung*
  Ziel dieses Vorhabens ist der Entwurf und die Entwicklung einer objektorientierten Laufzeitumgebung, auf deren Grundlage die Interaktionen zwischen Software-Komponenten hinsichtlich Laufzeiteffizienz optimiert werden können. In Abhängigkeit von dem gegebenen Anwendungsprofil sollen die jeweils aus Kostengesichtspunkten günstigsten *Interaktionsmuster* zum Einsatz kommen.

- *Interaktionsmethoden für die Informationsfusion*
  In diesem Projekt werden Methoden und Werkzeuge entwickelt, die es Benutzern ermöglichen, mit den zu fusionierenden bzw. den fusionierten Daten zu interagieren. Objektbewegungen werden dabei als ein besonders charakteristisches zu visualisierendes Merkmal im Fusionsprozeß betrachtet. Einerseits wird ein zusätzlicher Präsentationsvariablensatz benötigt, denn Visualisierungen der zu fusionierenden Daten verwenden oft für sich bereits die üblichen Variablen wie Farbe, Form und Position. Andererseits wird erwartet, daß es gute Entsprechungen für Fusionsprozesse in Form von visualisierten Objektbewegungen gibt.

- *Visuelles Data Mining zur Unterstützung der Informationsfusion*
  In diesem Forschungsvorhaben werden Methoden zur Datenvisualisierung und -manipulation entwickelt. Im einzelnen werden dabei betrachtet: Visualisierungstechniken für große Mengen heterogener Daten auf der Basis einer visuellen Verknüpfung von getrennten Visualisierungen der heterogenen Daten (z.B. durch *Linking and Brushing*), Techniken der Zoom-Visualisierung von großen Datenmengen durch (teilweise anwendungsbezogene) Verfahren zur visuellen Darstellung von Daten und Filter, die das jeweils wesentliche von Datensätzen anzeigen, die nicht im Fokus des Benutzers liegen und schließlich Techniken der interaktiven Manipulation.

- *Informationsgewinnung zur Unterstützung des Gußteilentwurfs*
  Zielstellung dieses Projektes ist die Optimierung und Rationalisierung des Prozesses der Rohteilgenerierung in Gießereibetrieben unter Verwendung verschiedener Informationsquellen. Um dies zu erreichen ist es notwendig, bereits vorhandene gießereispezifische Datenbanken mit dem Erfahrungswissen des Ingenieurs zu verbinden. Durch die Workbench sollen Abhängigkeiten zwischen den Daten der Datenbanken abgeleitet werden, um eine neue Qualität der Daten zu erreichen, so daß der beauftragte Mitarbeiter bei der Rohteilgenerierung unterstützt wird.

Das Gesamtprojekt wurde offiziell im Januar 2000 gestartet und hat eine Laufzeit von zunächst drei Jahren.

7 Zusammenfassung

Die Kombination von Daten aus unterschiedlichen Quellen sowie die darauf aufbauende Gewinnung von neuen Informationen durch Filterung, Verdichtung und Extraktion von Zusammenhängen ist eine Aufgabe, die in vielen Anwendungsgebieten besteht. Insbesondere auch zur Beherrschung des stetig zunehmenden Informationsaufkommens aufgrund des einfachen Zugriffs auf weltweit verfügbare Quellen wächst der Bedarf nach einer intelligenten Informationsfusion. Einen wichtigen Beitrag hierzu müssen Methoden und Techniken aus dem Datenbank- und Data Mining-Bereich leisten.
Im Rahmen der Forschergruppe „Workbench für die Informationsfusion" der Universität Magdeburg sollen wichtige Aspekte der beschriebenen Anforderungen hinsichtlich der Verbindung von Integrations-, Fusions- und Analysemethoden für heterogene, verteilte Datenbestände untersucht werden. Hierbei wird insbesondere angestrebt, das Fachwissen aus verschiedenen Bereichen gezielt zusammenzubringen, um so einerseits anwendungsspezifisches Wissen berücksichtigen zu können und andererseits die Anforderungen aus den verschiedenen beteiligten Bereichen effektiv und möglichst effizient erfüllen zu können.

Literatur

[AS94] R. Agrawal und R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proc. of the 20th Int. Conf. on Very Large Data Bases (VLDB)*, S. 478–499, Santiago, Chile, September 1994.

[AZ98] H. Arabnia und D. Zhu, Herausgeber. *Proc. of the Int. Conf. on Multisource-Multisensor Information Fusion - FUSION '98*, Las Vegas, NV, 1998. CSREA Press.

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

[CHY96] M. Chen, J. Han und P. Yu. Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):866–883, 1996.

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

[Fis95] D. Fisher. Optimization and simplification of hierarchical clustering. In *Proc. of 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD-95)*, S. 118–123, Montreal, Canada, August 1995.

[FPSS96] U. Fayyad, G. Piatetsky-Shapiro und P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth und R. Uthurusuamy, Herausgeber, *Advances in Knowlede Discovery and Data Mining*, Kapitel 1, S. 1–34. AAAI/MIT Press, Cambridge, MA, 1996.

[Ger98] M. Gertz. Managing Data Quality and Integrity in Federated Databases. In *2nd Annual IFIP TC-11 WG 11.5 Working Conf. on Integrity and Internal Control in Information Systems*, Warrenton, Virginia, November 1998. To appear.

[GGF+96] G. Gardarin, S. Gannouni, B. Finance, P. Fankhauser, W. Klas, D. Pastre, R. Legoff und A. Ramfos. IRO-DB — A Distributed System Federating Object and Relational Databases. In *Object-Oriented Multidatabase Systems — A Solution for Advanced Applications*, Kapitel 20, S. 684–712. Prentice Hall, Eaglewoods Cliffs, NJ, 1996.

[HCC92] J. Han, Y. Cai und N. Cercone. Knowledge Discovery in Databases: An Attribute-Oriented Approach. In *Proc. of 1992 Int. Conf. on Very Large Data Bases (VLDB'92)*, S. 547–559, Vancouver, Canada, August 1992.

[HHSS98] M. Höding, R. Hofestädt, G. Saake und U. Scholz. Schema Derivation for WWW Information Sources and their Integration with Databases in Bioinformatics. In *Advances in Databases and Information Systems – ADBIS'98, Poznań, Poland, September 1998*, LNCS 1475, S. 296–304, Berlin, 1998. Springer-Verlag.

[HKM+96] K. Hätönen, M. Klemettinen, H. Mannila, P. Ronkainen und H. Toivonen. Knowledge Discovery from Telecommunication Network Alarm Databases. In *Proc. of 12th Int. Conf. on Data Engineering (ICDE'96)*, S. 115–122, New Orleans, 1996.

[HKW99] A. Hinneburg, D.A. Keim und M. Wawryniuk. HD-Eye: Visual Mining High-dimensional Data. *IEEE Computer Graphics and Applications*, 1999.

[HMPU97] D. Heckerman, H. Mannila, D. Pregibon und R. Uthurusamy, Herausgeber. *KDD-97 – Proc. of the 3rd Int. Conf. on Knowledge Discovery and Data Mining*, Menlo Park, CA, 1997. AAAI Press.

[Inm96] W. H. Inmon. *Building the Data Warehouse*. Wiley & Sons, 2 Auflage, 1996.

[JW93] G. Jakobson und M.D. Weissman. Alarm Correlation. *IEEE Network*, 7(6):52–59, November 1993.

[Kei00] D.A. Keim. Designing Pixel-oriented Visualization Techniques: Theory and Applications. *IEEE Transactions on Visualization and Computer Graphics*, 2000.

[KK96] D. Keim und H.-P. Kriegel. Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):923–938, December 1996.

[LA94] Y. Leung und M. Apperley. A Review and Taxonomy of Distortion-oriented Presentation Techniques. In *Proc. Human Factors in Computing Systems*, S. 126–160, Boston, MA, 1994.

[LK95] R.C. Luo und M.G. Kay, Herausgeber. *Multisensor Integration and Fusion for Intelligent Machines and Systems*. Ablex Publishing Corporation, Norwood, NJ, 1995.

[Mat97] R. Mattison. *Data Warehousing and Data Mining for Telecommunications*. Artech House, Norwood, MA, 1997.

[MTV95] H. Mannila, H. Toivonen und A.I. Verkano. Discovering frequent episodes in sequences. In *Proc. of 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD-95)*, S. 210–215, Montreal, Canada, August 1995.

[MY95] W. Meng und C. Yu. Query Processing in Multidatabase Systems. In W. Kim, Herausgeber, *Modern Database Systems*, S. 551–572, New York, NJ, 1995. ACM Press.

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

[PGV95] S. Pfleger, J. Goncalves und D. Vernon, Herausgeber. *Data Fusion Applications*. Springer-Verlag, Berlin, 1995. Research Reports ESPRIT.

[SHF96] E. Simoudis, J. Han und U. Fayyad, Herausgeber. *KDD-96 – Proc. of the 2nd Int. Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA, 1996. AAAI Press.

[Shn96] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Conference on Visual Languages*, S. 336–343, 1996.

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

[War94] M.O. Ward. XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data. In *Visualization 94*, S. 326–336, Washington, DC, 1994.

[Wro98] S. Wrobel. Data Mining und Wissensentdeckung in Datenbanken. *Künstliche Intelligenz – Organ des FB 1 der Gesellschaft für Informatik (GI)*, (1), 1998.

[WWSE96] S. Wrobel, D. Wettschereck, E. Sommer und W. Emde. Extensibility in data mining systems. In Simoudis et al. [SHF96].

[YPAGM98] R. Yerneni, Y. Papakonstantinou, S. Abiteboul und H. Garcia-Molina. Fusion queries over internet databases. In *Advances in Database Technology - EDBT'98*, LNCS 1377, S. 57–71. Springer-Verlag, 1998.