Visuelle Exploration und semantikbasierte Fusion multivariater Datenbestände

Stefan Audersch, Guntram Flach, Tom Klipps
Zentrum für Graphische Datenverarbeitung e.V., Rostock
Joachim-Jungius-Str. 11, 18059 Rostock
{stefan.audersch, guntram.flach, tom.klipps}@rostock.zgdv.de

Abstract: In verschiedenen Forschungsbereichen spielen Techniken der semantischen Datenintegration eine besondere Rolle. Es existiert ein Bedarf an Lösungen, die über eine Fusion von Daten hinaus eine Nutzung verteilt entwickelter Analysemethoden für Daten ermöglichen.
In dieser Arbeit wird ein Ansatz entwickelt, der basierend auf Techniken der Datenexploration und semantikbasierten Fusion eine Nutzung von Analysemethoden wie DataMining- und Visualisierungstechniken in verteilten Umgebungen erlaubt. Unter Einsatz von Ontologien zur semantischen Beschreibung verteilter Quellen wird es ermöglicht, die Daten und Analysemethoden aus diesen Quellen zu fusionieren.
Es wird eine Architektur für diese Aufgabe vorgestellt. Kern der Architektur ist die Gatewaykomponente, die es dem Analysten erlaubt, Daten und Analysemethoden in einer verteilten Umgebung zu nutzen. Ein Prototyp implementiert die vorgestellten Komponenten.

1 Einleitung

Die automatische Erfassung von Daten durch kommerzielle Geräte und wissenschaftliche Instrumente führen zu immer größeren Mengen von immer komplexeren Daten, deren manuelle Analyse die kognitiven Fähigkeiten des Analysten bei weitem überschreiten. Zur Automatisierung dieser Analysen kommen Techniken aus dem Bereich des Knowledge Discovery in Databases (KDD) zum Einsatz, bei dem Data Mining und Visualisierung zentrale Schritte darstellen. Die visuelle Datenexploration erlaubt es dem Benutzer, einen schnellen Einblick in die Struktur der Daten zu bekommen, Schlussfolgerungen aus den Daten zu ziehen sowie direkt mit den Daten zu interagieren (Overview, Zoom and filter, details-on-demand) [AS04]. Die Qualität der Ergebnisse einer solchen Wissensgewinnung ist stark von dem Expertenwissen abhängig, mit dessen Hilfe die eingesetzten Verfahren gesteuert werden. Neben dem benötigten Wissen ist eventuell die Datengewinnung, Vorverarbeitung bzw. Aufbereitung von Daten nur in einer speziellen Laborumgebung oder unter Einsatz besonderer Mittel, Werkzeuge oder Analysemethoden möglich. Besonders auf den Gebieten der Medizin und Molekularbiologie führt die thematische und räumliche Trennung der weltweiten Forschung dazu, dass eine Vielzahl von Firmen, Gruppen und Konsortien existieren, von denen jede ihre eigene Basis an Forschungsdaten besitzt. Eine Unterstützung der Forschungsarbeit können Werkzeuge und Verfahren bieten, welche die Daten der durchgeführten Experimente mit Informationen aus komplementären Datenquellen anreichern und eine Einordnung und Bewertung der eigenen Daten im Vergleich mit Daten anderer Forschungen ermöglichen. Dabei ergibt sich die Notwendigkeit einer dynamischen Informationsfusion, die eine bedarfsgetriebene, skalierbare Kopplung und Integration von Datenbanken, Datenströmen und Datenanalysemodellen verwirklicht.
Ausgangspunkt dieses Beitrages ist ein Anwendungsszenario, das medizinische Messwerte im Rahmen einer Klinischen Studie[1] betrachtet. In diesem Szenario geht es um die Analyse multivariater Patientendaten (z.B. Nerven-, Leber- und Blutdaten) unter Einsatz von Data Mining und Visualisierungstechniken in verteilten Umgebungen. Obwohl die erhobenen Daten eine semantische Einheit bilden, werden sie aufgrund unterschiedlicher technischer und fachlicher Anforderungen in Teildatenbestände zerlegt und getrennt voneinander analysiert.

[1] In Kooperation mit der Teraklin AG (http://www.teraklin.de)

2 Problemstellung und Anforderungen

Die getrennten Datenbestände werden einzeln ausgewertet, aufbereitet und analysiert. Hierbei kommen unterschiedliche DataMining- und Visualisierungstechniken zum Einsatz, die auf die Beschaffenheit der unterschiedlichen Daten zugeschnitten sind (Abbildung 1). Der Zusammenhang zwischen den einzelnen Daten kann in dieser Phase der Wissensgewinnung nicht erfasst werden. Erst eine zentrale Anwendung, die Zugriff auf die einzelnen Datenbestände (Explorationsquellen) hat, leistet dies.

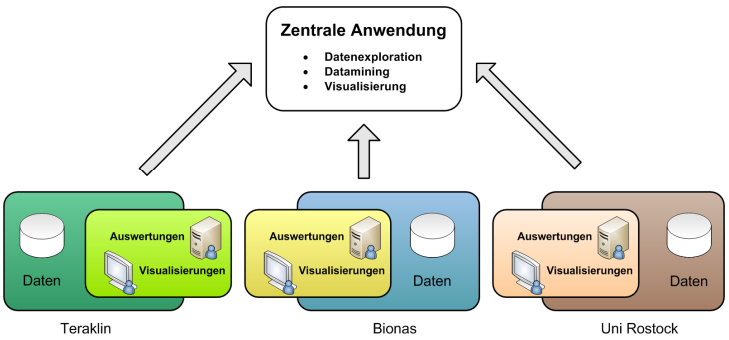

Abbildung 1: Anwendungsszenario

Eine Lösung in diesem Sinne würde es erlauben, auf der gesamten Datenbasis weiterzuarbeiten, wodurch sich beispielsweise die beiden folgenden Aufgabenstellungen lösen ließen:

- Die Visualisierung von Zusammenhängen zwischen Komazustand, Blut- und Leberwerten der Patienten ist auf Grundlage der Teraklin-Datenbasis möglich. Im Rahmen der Studie ist es aber durchaus von Interesse, Messungen an Gewebeproben (Bionas) oder Werte von Aminosäuren (Universität) in einen Zusammenhang mit Blut-, Leber- oder Komawerten zu bringen.
  Eine Kombination von Rohdaten und Ergebnissen einer Datenbasis mit Rohdaten und Ergebnissen anderer Datenbasen ist notwendig, um diese Analysen zu realisieren.
- Zwischen unterschiedlichen Datenreihen der Teraklin-Datenbasis wird ein Zusammenhang vermutet, der sich anhand der vorliegenden Daten nicht eindeutig zeigen lässt. In den Auswertungen der Bionas- oder Universitäts-Daten wird festgestellt, dass ein bestimmtes Phänomen bei einem Teil der Patienten auftritt. Denkbar ist an dieser Stelle die Durchführung der ursprünglichen Analyse der Teraklin-Daten auf dem extern motivierten Teilbereich. Es muss möglich sein, selektive Anfragen an Daten einer Datenbasis unter Ausnutzung von Inhalten anderer Datenbasen zu formulieren. Das Formulieren beliebiger Anfragen an einen virtuellen Gesamtdatenbestand wäre in diesem Fall die optimale Lösung.

Zur Lösung dieser Aufgaben besteht die Notwendigkeit, die Daten sowie die Analyseprozesse aus den verschiedenen Explorationsquellen virtuell zu fusionieren. Voraussetzung für eine intelligente Zusammenführung ist eine maschinenverständliche Semantik der Explorationsquellen. Grundlage hierfür bietet eine gemeinsame Ontologie, die unter anderem eine gemeinsame Terminologie abbildet. Die angedachten Anforderungen sollen nachstehend zusammengefasst und konkretisiert werden:

- **Datenintegration**: Es soll möglich sein, auf der gesamten Datenbasis zu arbeiten, ohne die Daten in einem initialen Schritt in eine einzige Datenbasis zu integrieren.
- **Datenexploration**: In der Datenexploration soll es möglich sein, den gesamten Datenbestand sowie auch die Ergebnisse der lokal durchgeführten Analysen, DataMining-Verfahren und Visualisierungen einzusehen.
- **DataMining und Visualisierung**: Es soll möglich sein, vorhandene DataMining-Verfahren oder Visualisierungen auf neuen (aus der Datenfusion resultierenden) Datenbeständen durchzuführen. Im Sinne einer visuellen Datenexploration soll eine Interaktion mit bestimmten Visualisierungen möglich sein.

- **Semantik**: Die semantischen Beschreibungen sollen es erlauben, verschiedene Datenquellen einfach miteinander zu verbinden und deren Heterogenität aufzulösen. Durch Nutzung der Semantik sollte das Anwenderprogramm dem Benutzer Hilfestellung (z.B. in Form eines Wizards) bei der Exploration geben.

3 Realisierungsaspekte

Die für den Lösungsansatz notwendigen Überlegungen werden im folgenden Abschnitt durch eine Auswahl verschiedener Realisierungsaspekte kurz vorgestellt.

Prozesse

Verfahren des Data Mining und der Visualisierung von Daten stellen zentrale Schritte im KDD-Prozess dar und bilden die Grundlage für die visuelle Datenexploration. Um diese Verfahren in das System zu integrieren, können diese als Prozesse aufgefasst und als Service von einer Explorationsquelle bereitgestellt werden. Ebenso lässt sich die Bereitstellung von Datentabellen als auch der Zugriff auf Analyseergebnisse als Prozess definieren. Eine Explorationsquelle (Abbildung 2) kann verschiedene Prozesse zur Verfügung stellen. Für die Integration von Prozessen ist es notwendig, die Explorationsquellen und deren Prozesse semantisch zu beschreiben. Die Beschreibungen umfassen dabei Informationen über die von der Explorationsquelle bereitgestellten Prozesse. Für den jeweiligen Prozess sind Informationen über dessen Vorbedingungen, Parameter und Ergebnisse definiert.

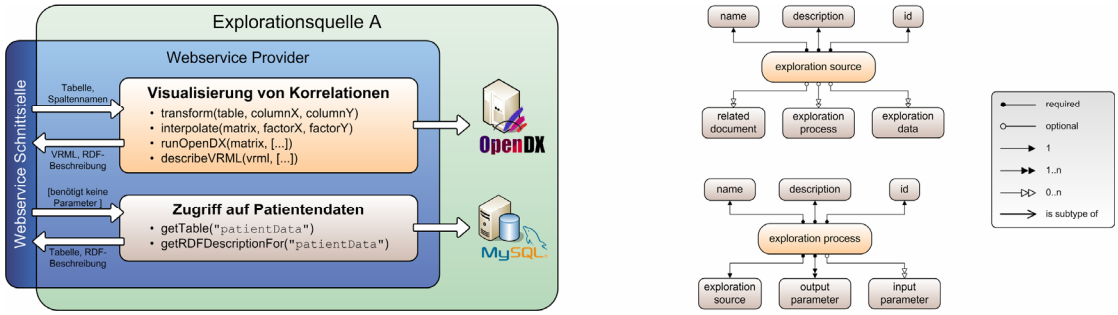

Abbildung 2: Explorationsquelle und semantische Beschreibungen

Im Rahmen der Arbeiten wurden verschiedene Prozesse der Systeme Weka (Data Mining), OpenDX (Visualisierung) und JFreeChart (Visualisierung) entsprechend semantisch beschrieben und als Web Service zur Verfügung gestellt.

Semantische Daten- und Prozessintegration

Auf der Grundlage der semantischen Beschreibung kann die Integration der Daten und Prozesse erfolgen. Bei der Integration von Datentabellen kann hierdurch von Tabellen- und Attributnamen abstrahiert werden [AF04]. Existiert beispielsweise in einer Explorationsquelle ein Prozess P1, der die Korrelation eines Blutwertes zu einem Leberwerte (HE in T1) visualisiert, so kann dieser Prozess nun auch zur Darstellung der Korrelation zu einem anderen Leberwert (MELD in T2) aus einer anderen Explorationsquelle verwendet werden (Abbildung 3).

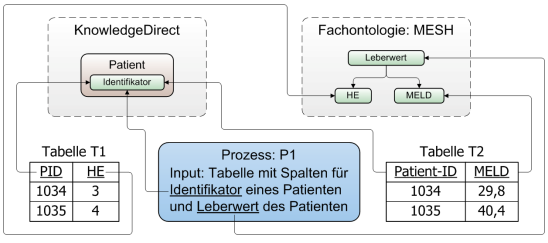

Abbildung 3: Nutzung semantischer Informationen

Auf der Grundlage der semantischen Beschreibungen ist es ebenfalls möglich, Datenbestände zusammenzuführen (Semantic Join) [LR03] und auf deren Basis neue Visualisierungen bzw. Data Mining-Verfahren anzuwenden. So kann beispielsweise mit einem geeigneten Prozess der Zusammenhang zwischen den beiden Leberwerten HE und MELD (Abbildung 3), die sich in unterschiedlichen Explorationsquellen befinden und über den Patientenidentifikator verbinden lassen, dargestellt werden.

Hilfestellungen durch semantische Informationen

Die semantischen Beschreibungen bieten durch Hilfestellungen oder semantische Kontrollen ebenfalls Potential für die Unterstützung des Benutzers bei der Exploration. Visualisierungs- und Data Mining Prozesse lassen sich hinsichtlich Ihrer Eignung für bestimmte Datenstrukturen beschreiben. Auf Basis der im System vorhandenen Metadaten zu den verschiedenen Datentabellen können Vorschläge zur Eignung der bereitgestellten Prozesse gemachen werden [NS04]. So bieten sich beispielsweise Visualisierungen wie Shape Coding oder Parallele Koordinaten erst bei einer großeren Anzahl von Attributen an.
Durch die Nutzung von Fachontologien lassen sich die in den Explorationsquellen bereitgestellten Analyseergebnisse semantisch einordnen und somit besser für weitere Recherchen nutzen. Zudem bieten die Fachontologien dem Benutzer Hilfestellung bei der Auswahl von Attributen. So kann beispielsweise für einen Prozess die Auswahl von Leberwerten (HE, MELD) über die in der Fachontologie enthaltenen Beziehungen leicht erfolgen.

4 Architektur und prototypische Implementierung

Die entwickelten Konzepte wurden in der prototypischen Implementierung *KnowledgeDirect* [Kl04] umgesetzt. Zentraler Kern der Architektur ist das *Knowledge Explore Gateway*, bestehend aus *Control*, *Retrieval* und der *Integration Engine*. Aufgabe der Integration Engine ist die Einbindung verschiedener Explorationsquellen, welche mit der semantischen Integration der in den Quellen bereitgestellten Datenstrukturen, Analyseergebnissen, Analyseprozessen, Data Mining und Visualisierungstechniken einhergeht.

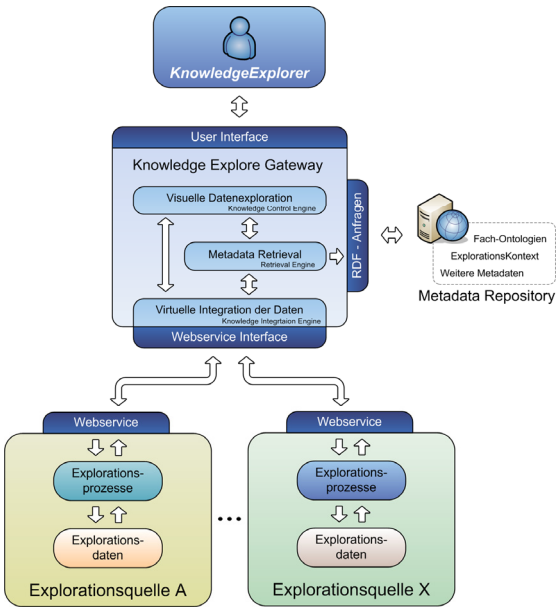

Abbildung 4: KnowledgeDirect-Architektur

Grundlage für die Integration sind semantische Beschreibungen auf der Basis von RDF und OWL, die im Metadatenrepository verwaltet werden. Einen adäquaten Zugriff auf die Metadaten erhält das Knowledge Explore Gateway über die Retrieval Engine (RQL). Die Knowledge Control Engine ermöglicht die einfache Kombination der einzelnen Funktionen und Daten der Explorationsquellen auf Basis der semantischen Beschreibungen und erlaubt weiterhin die Erweiterung um komplexe

Funktionalität auf dem Gebiet des Data Minings und der Visualisierung, wie z.B. 3D-Darstellungen. Der Zugriff auf verschiedene Quellen erfolgt auf der Basis von Web Services und derart transparent, dass sämtliche Aktionen explorationsübergreifend möglich sind. Eine Zuordnung von Daten zu einer bestimmten Explorationsquelle dient nur der Orientierung und birgt keine Einschränkungen in Bezug auf die Nutzung dieser Daten im Zusammenhang mit anderen Explorationsquellen.

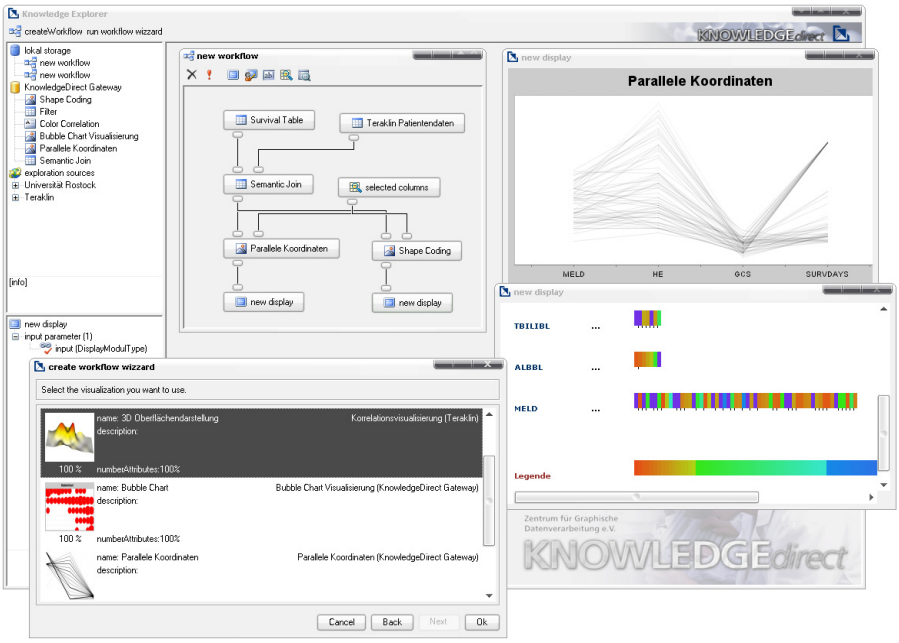

Abbildung 5: KnowledgeDirect Benutzerschnittstelle

Im Rahmen einer prototypischen Implementierung wurden die entwickelten Konzepte und Methoden innerhalb des medizinischen Anwendungsszenarios evaluiert.

5 Zusammenfassung und Ausblick

Das entwickelte KnowledgeDirect-Framework dient der universellen Exploration und semantikbasierten Fusion multimedialer sowie multivariater Datenbestände und ermöglicht es, die getrennt ermittelten Analyseergebnisse unter Nutzung von Ontologiewissen zusammenzuführen. Die von den Explorationsquellen bereitgestellten parametrisierbaren Data Mining- und Visualisierungstechniken, sowie die Analyseprozesse, Rohdaten und aggregierten Daten lassen sich integrieren und erlauben, globale Analysen über den gesamten Datenbestand durchzuführen. Mit semantisch gesteuerter Interaktions- und Navigationstechnik wird es auf einfache Weise ermöglicht, Daten aus verschiedenen Explorationsquellen zu selektieren, zu kombinieren, anzuzeigen und mit ihnen zu interagieren.

Literaturverzeichnis

[AF04] S. Audersch, G. Flach: Universeller Gateway-Ansatz auf der Basis semantisch angereicherter Web Services im Rahmen heterogener eGovernment-Anwendungen, 16. Workshop über Grundlagen von Datenbanken, Monheim, 2004.

[AS04] C. Ahlberg, B. Shneiderman: Visual Information Seeking: Tight coupling of dynamic query filters with starfield displays. Proceedings of ACM CHI94, S.313-317, 1994

[Kl04] T. Klipps: Exploration und semantikbasierte Fusion multivariater Datenbestände in domänenspezifischen Anwendungsumgebungen. Universität Rostock, Diplomarbeit, 2004.

[LR03] U. Leser, P. Rieger: Integration molekularbiologischer Daten. Datenbank-Spektrum 3 (2003) Nr. 6, S. 56-66.

[NS04] T. Nocke, H. Schumann: Meta Data for Visual Data Mining. Proceedings Computer Graphics and Imaging, CGIM 2002, Kaua'i, Hawaii, USA, 2002.