Ad-hoc-Integration in schemabasierten P2P-Systemen

Marcel Karnstedt[1] Kai-Uwe Sattler[1] Eike Schallehn[2] Martin Endig[3]
[1]Technische Universität Ilmenau, Postfach 100565, 98684 Ilmenau
[2]Universität Magdeburg, Postfach 4120, 39016 Magdeburg
[3]Fraunhofer-Institut IFF.FHG, Sandtorstr. 22, 39106 Magdeburg

Abstract: Die weite Verfügbarkeit von Daten zu verschiedensten Aspekten in elektronisch verarbeitbarer Form eröffnet für Anwendungen wie das Katastrophenmanagement neue Möglichkeiten der entscheidungsunterstützenden Informationsbereitstellung. Notwendig ist dafür jedoch eine dynamische, aufgabengetriebene Verbindung der verschiedenen Datenbestände, die durch klassische Datenintegrationstechniken nur bedingt erzielt werden kann. In diesem Beitrag diskutieren wir daher Techniken, die eine derartige Ad-hoc-Integration auf der Basis von schemabasierten Peer-to-Peer (P2P)-Systemen unterstützen. Im Mittelpunkt stehen dabei die Probleme der Korrespondenzfindung und -definition sowie eine effiziente Anfrageverarbeitung.

1 Einleitung

Das Problem der Integration heterogener Datenbestände ist seit vielen Jahren Gegenstand aktiver Forschung im Datenbankbereich. Neben Data-Warehouse-Ansätzen, die auf der Extraktion von Daten aus den Quellsystemen und deren Materialisierung in Form eines integrierten Bestandes basieren, wurden u.a. auch Techniken für heterogene Datenbanksysteme entwickelt, die einen virtuellen Integrationsansatz verfolgen. Derartige Lösungen haben sich insbesondere für die Integration von Web-Quellen mit Hilfe von Mediatorsystemen etabliert. Dennoch stoßen diese klassischen „Integrationstechniken“ in bestimmten Anwendungsbereichen an Grenzen.

Als Anwendungsfall betrachten wir hierbei die Integration von Informationssystemen in Katastrophenszenarien, wie sie zum Beispiel im Fall der Flut entlang der Elbe und ihrer Nebenflüsse im Sommer 2002 von Vorteil gewesen wäre. Bei einem derartigen Ereignis werden relevante Daten von lokalen und übergreifenden Krisenstäben sowie den verschiedenen an Einsätzen beteiligten Organisationen, z.B. Technisches Hilfswerk oder Feuerwehren, sowohl bereitgestellt als auch zur Koordinierung des gemeinsamen Vorgehens benötigt. Hierbei handelt es sich teilweise um relativ statische Daten, wie z.B. Geo-Daten oder personenbezogene Daten. Andererseits sind für eine erfolgreiche Koordination auch vergleichsweise dynamische Daten notwendig, wie aktuelle Pegelstände der betroffenen Flüsse oder Füllstände in Stausystemen. Sollen diese Informationen im Katastrophenfall für die oben beschriebenen Nutzer zugänglich gemacht werden, ist insbesondere der voraussichtlich knappe Zeitrahmen ein restriktives Kriterium. Ebenso ist die Integration stark durch den aktuellen Informationsbedarf der Nutzer getrieben, weshalb unter Umständen

im laufenden Betrieb eine Anzahl von neuen Datenquellen hinzugezogen werden muss. Derartige Anwendungen sind somit gekennzeichnet durch eine große Anzahl lose gekoppelter Quellsysteme. Das primäre Ziel ist dabei nicht die vollständige Integration, wie beispielsweise in klassischen Schemaintegrationstechniken, sondern ein kurzfristiger zielgetriebener Zugriff auf verteilt vorliegende heterogene Datenbestände.

Für die Ad-hoc-Integration in einem derartigen Anwendungsfall gehen wir von folgenden Annahmen aus: *(1)* Es existiert eine gemeinsame Domäne für die *relevanten* Quellen. *(2)* Bilaterale „Absprachen“ im Sinne der Festlegung von Korrespondenzen zwischen den zu integrierenden Systemen sind einfacher zu treffen als die Durchführung einer vollständigen Schemaintegration. *(3)* Alle Quellsysteme verfügen über Mechanismen zur verteilten Anfrageverarbeitung bzw. können durch Wrapper um solche Funktionalitäten erweitert werden. Im Weiteren werden wir hierzu einen P2P-basierten Ansatz vorstellen und dabei insbesondere die Schritte der Korrespondenzdefinition sowie der Anfrageverarbeitung detaillierter betrachten.

2 Schemabasierte P2P-Systeme

Eine geeignete Infrastruktur für den oben beschriebenen Kontext bilden schemabasierte P2P-Systeme [TIM+03], wobei die Quellen die im Wesentlichen gleichberechtigten Peers darstellen. Solche Systeme kennzeichnen sich zum einen dadurch, dass alle Peers über die gleichen Anfragemöglichkeiten verfügen, eine Anfrage über dem Gesamtdatenbestand also an jedem Peer gestellt werden kann. Desweiteren ist es charakteristisch, dass kein Peer über vollständiges globales Wissen wie Verteilungsschema, beteiligte Peers etc. verfügt. Demzufolge kann über die Verarbeitung bzw. Weiterleitung von (Teil-)Anfragen nur lokal für jeden beteiligten Peer entschieden werden. Daraus ergeben sich eine Reihe von Vorteilen: *(1)* eine potentiell höhere Skalierbarkeit und Robustheit aufgrund des Fehlens einer zentralen Koordinationsinstanz sowie *(2)* die Möglichkeit bereits verfügbare Korrespondenzen zu nutzen, so dass ein einzelnes Peer durch die Definition einer Korrespondenzbeziehung zu einem bereits teilnehmenden Peer sofort auf alle Daten des Systems zugreifen kann. Letzteres wird möglich, indem Anfragen des neuen Peers über andere Peers weitergeleitet werden oder indem aus den existierenden Korrespondenzen neue (transitive) Korrespondenzen zu den relevanten Peers abgeleitet werden [MH03].

Unter Verwendung einer P2P-Infrastruktur kann eine Ad-hoc-Integration wie folgt realisiert werden. Den Ausgangspunkt bilden dabei eine gemeinsame Ontologie, die grundlegende Konzepte und deren semantische Beziehungen für die Integrationszieldomäne definiert, sowie für jedes Peer eine Menge erreichbarer Peers bzw. die Möglichkeit solche zu identifizieren (z.B. in Form eines Directory-Dienstes):

1. Peers versuchen, zu einem oder mehreren anderen Peers Schema-Korrespondenzen aufzubauen. Hierfür sind insbesondere Schema-Matching-Techniken [RB01] geeignet, die in Verbindung mit domänenspezifischen Matching-Regeln eine (semi-)automatische Ableitung der Korrespondenzen ermöglichen. Das sich daraus ergebende Korrespondenz-Netz kann entweder statisch (paarweise Korrespondenzen werden

vorab definiert) oder dynamisch (neue Korrespondenzen werden im Verlauf der Anfrageverarbeitung identifiziert) sein.

2. Mit den Korrespondenzen werden auch die für das Anfrage-Routing benötigten Informationen gesammelt, die in Form von Routing-Indexen eine effiziente Anfragezerlegung bzw. -weiterleitung ermöglichen.
3. Anschließend können Anfragen formuliert und ausgeführt werden, die ggf. die Etablierung weiterer Korrespondenzen initiieren.

Es sei angemerkt, dass ein derartiger Integrationsansatz natürlich auch Einschränkungen aufweist. Aufgrund der losen Koppelung und dem Verzicht auf ein integriertes globales Schema mit direkten Korrespondenzen zu allen Quellen können Korrektheit und Vollständigkeit der Ergebnisse nicht mehr uneingeschränkt garantiert werden. Besondere Herausforderungen stellen somit der Umgang mit unscharfen, ähnlichkeitsbasierten Anfragen sowie die Bewertung bzw. Vorhersage der Qualität von Anfrageergebnissen dar. Auf einige dieser Aspekte werden wir im Abschnitt 5 kurz eingehen.

3 Schema Matching

In dem von uns beschriebenen Szenario stellen alle integrierten Quellen ihre Daten in Form von XML bereit, sowie das zugehörige DTD oder XML-Schema. Um die Beziehungen zwischen den Quellsystemen zu formulieren werden Korrespondenz-Operationen auf der Basis von Äquivalenz bzw. Subelement-Beziehungen eingeführt. Für genauere Ausführungen zu den verwendeten Operationen verweisen wir auf [KHS04]. Anfragen werden über einer Teilmenge von XQuery formuliert, korrespondierend zu XPath mit Verbunden. Wie schon in Abschnitt 2 eingeführt, bietet sich für die Definition der Peer-Korrespondenzen der Einsatz von Schema-Matching-Techniken an. Schema Matching behandelt das Problem der Ableitung einer Abbildung zwischen Elementen zweier Schemata, die semantisch zueinander korrespondieren [RB01]. Hier bietet sich die Verwendung einer Ontologie als Beschreibung der Domäne des Integrationsziels an. Diese kann zur Unterstützung einzelner Matcher verwendet werden, dass Matching findet also direkt zwischen zwei Peer-Schemata statt. In einer zweiten Variante führt jedes Peer ein Matching seines Schemas mit der Ontologie durch und die Korrespondenzen zwischen Peers werden daraus abgeleitet.

Auch unter Verwendung von Schema-Matching-Techniken kann jedoch die Integration im allgemeinen Fall nicht vollautomatisch durchgeführt werden – zu vielfältig sind die möglichen strukturellen und semantischen Heterogenitäten. In [SSC04] haben wir eine regelbasierte Erweiterung für Schema-Matching-Ansätze vorgestellt, die eine einfache Einbeziehung domänenspezifischer Matcher ermöglicht. Ausgehend von atomaren Matchern für einfache Attribute in Form von Prädikaten können dabei komplexe domänenspezifische Matcher für Klassen bzw. Relationenschemata in Form von Regeln formuliert werden. Im folgenden Beispiel wird dies illustriert, indem atomare Matcher für Elementnamen (match_name) auf der Basis von Gleichheit oder der Editierdistanz zunächst zu Matchern für Attribute (match_attribute) kombiniert werden, die wiederum Namen und Typ vergleichen. Schließlich können daraus Klassen-Matcher erstellt werden, die über der Attributmenge arbeiten (hier nur angedeutet durch das Prädikat match_children):

match_name(n_1, n_2, 1.0) $:-$ equal(n_1, n_2).
match_name(n_1, n_2, *conf*) $:-$ edistance(n_1, n_2, *conf*).
match_attribute(a_1, a_2, *conf*) $:-$ match_name(a_1.name,a_2.name,*conf*$_1$),
match_type(a_1.type, a_2.type, *conf*$_2$), *conf* $= f$(*conf*$_1$, *conf*$_2$).
match_classes(c_1, c_2, *conf*) $:-$ match_name(c_1.name,c_2.name, *conf*$_1$),
match_children(c_1.attribs, c_2.attribs, *conf*$_2$), *conf* $= f$(*conf*$_1$, *conf*$_2$).

Eine solche Regelbasis kann als Teil der Ontologie für das Matching bereitgestellt werden und somit von allen Peers (bzw. den Integratoren dieser Peers) genutzt werden. Das Hauptproblem besteht hier im Finden und Evaluieren der verwendeten Regeln. Der zweite Schritt nach der Identifikation korrespondierender Schema-Elemente ist die Ableitung der Mappings, die für Anfragedekomposition und -übersetzung benötigt werden, entweder direkt anhand der Regeln oder unter Zuhilfenahme der definierten Mappings und der verwendeten Ontologie.

4 Anfrageverarbeitung

In schemabasierten P2P-Systemen muss jedes Peer Anfragen unter Verwendung von unvollständigem Wissen bearbeiten. Zur Verfügung stehen nur die lokalen Daten und ihr Schema, sowie die Korrespondenzen zu den direkt benachbarten Peers. Es existieren keine zentralen Instanzen oder Mechanismen, da diese generell das P2P-Paradigma verletzen würden. Die aus verteilten Systemen bekannten Anfragestrategien *Data Shipping* und *Query Shipping* können nicht ohne Modifikationen effizient auf P2P-Systeme übertragen werden. Ein entsprechender Ansatz wird zum Beispiel in [PM02] beschrieben.

Kann ein Peer eine Anfrage lokal nicht (vollständig) beantworten, so benötigt es Informationen über die Qualität der Daten, die über die Verbindungen zu den benachbarten Peers angefragt werden können. Eine Möglichkeit ist hier die Verwendung einer Form von Routing-Tabellen, sogenannten *Routing-Indexen*. In unserem Szenario verwenden wir eine Form von *Hop Count Routing Indices* [CGM02]. Im Gegensatz zu den dort beschriebenen Indexen erfassen wir neben der Instanzebene, beschrieben durch Filterprädikate, auch die Schemaebene in den Indexeinträgen. Auf nähere Ausführungen zu den verwendeten Indexen wird aus Platzgründen verzichtet. Erste Erfahrungen, Ergebnisse und Folgerungen aus der Verwendung der Indexe beschreiben wir unter anderem in [KHS04]. Das Schema eines Indexeintrages soll hier zur Veranschaulichung genügen:

(Nachbar ID, Kategorie (Schemaebene), Prädikat (Instanzebene), Kardinalität, Anzahl Peers)

In [CGM02] erfolgt der Aufbau der Indexe durch Austausch entsprechender Aggregationsnachrichten zwischen den Peers. Wir können dies mit der Bestimmung der bidirektionalen Korrespondenzen zwischen den Quellsystemen verbinden. Dabei kann ein Peer bei Eintritt in das Netzwerk die benötigten Informationen beim kontaktierten Peer anfragen, wobei dies mittels der Techniken der Anfrageverarbeitung die Daten von weiteren Peers erhalten und kombinieren kann. Alternativ können die Informationen auch aus dem lokalen Index extrahiert werden, wobei der vorgegebene Horizont sowie die Kohärenz der Daten Beachtung finden müssen.

5 Ausblick

Ad-hoc-Integration heterogener Datenbestände ist eine Aufgabenstellung, die, aufgrund der Vielzahl potentiell nützlicher Datenquellen, deren Verteilung und Heterogenität, mit neuen Herausforderungen an Datenintegrationstechniken verbunden ist. Ein möglicher Ansatz hierfür ist die Verteilung und Dezentralisierung der eigentlichen Integrationsarbeiten durch den Einsatz einer schemabasierten P2P-Infrastruktur, so dass *(1)* nur lokal beschränkte bzw. bilaterale Abstimmungen erforderlich sind und *(2)* sich aus den lokalen Interaktionen (in Form von Korrespondenzen sowie Anfrageverarbeitung) eine globale Gesamtsicht für das jeweilige Integrationsziel ergibt.

In diesem Beitrag haben wir erste Arbeiten zu dieser Problemstellung vorgestellt. Eine Reihe weiterer Aufgaben ist für die Zukunft vorgesehen, so u.a. die Behandlung von Datenkonflikten im Rahmen der Anfrageverarbeitung bei der dargestellten P2P-Integration. Besonders bei identifizierenden Attributen, die auch für die Anfrageverarbeitung benutzt werden, ist mit abweichenden Darstellungen durch fehlerhafte Eingaben oder unterschiedliche Eingabekonventionen zu rechnen. In materialisierten oder mediatorbasierten Integrationsszenarien können hierfür ähnlichkeitsbasierte Operationen angewandt werden, wie wir es zum Beispiel in [SSS04] dargestellt haben. Diese Ansätze können teilweise auf das in diesem Papier dargestellte Szenario übertragen werden und davon ausgehend kann zum Beispiel eine sukzessive Anfrageerweiterung bei der verteilten Bearbeitung umgesetzt werden. Desweiteren untersuchen wir derzeit Aspekte der Qualität von Anfrageergebnissen und arbeiten an einem dynamischen Kostenmodell.

Literatur

[CGM02] Crespo, A. und Garcia-Molina, H.: Routing indices for peer-to-peer systems. In: *Proc. of the 28th Conference on Distributed Computing Systems*. July 2002.

[KHS04] Karnstedt, M., Hose, K., und Sattler, K.-U.: Query Routing and Processing in Schema-Based P2P Systems. In: *Proc. DEXA 2004 (Workshop GLOBE'04), To appear*. 2004.

[MH03] Madhavan, J. und Halevy, A. Y.: Composing Mappings Among Data Sources. In: *VLDB 2003, Berlin*. S. 572–583. 2003.

[PM02] Papadimos, V. und Maier, D.: Mutant Query Plans. *Information and Software Technology*. 44(4):197–206. April 2002.

[RB01] Rahm, E. und Bernstein, P.: A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*. 10(4):334–350. 2001.

[SSC04] Saake, G., Sattler, K., und Conrad, S.: Rule-based Schema Matching for Ontology-based Mediators. *Journal of Applied Logic*. 2004. *To appear*.

[SSS04] Schallehn, E., Sattler, K., und Saake, G.: Efficient Similarity-based Operations for Data Integration. *Data and Knowledge Engineering Journal*. 48(3):361–387. 2004.

[TIM+03] Tatarinov, I., Ives, Z., Madhavan, J., Halevy, A., Suciu, D., Dalvi, N., Dong, X., Kadiyska, Y., Miklau, G., und Mork, P.: The Piazza Peer Data Management Project. *SIGMOD Record*. 32(3):47–52. 2003.