Qualitäts- und Semantik-gesteuerte Anfragebearbeitung für Peer-basierte Datenmanagementsysteme (PDMS)

Armin Roth
DaimlerChrysler Forschungszentrum Ulm
armin.roth@daimlerchrysler.com

Felix Naumann
Humboldt-Universität zu Berlin
naumann@informatik.hu-berlin.de

Abstract: Integrierte Informationssysteme basieren meist auf einem globalen Schema, dessen Bildung und Wartung aufwändig ist. Praktiker bevorzugen jedoch den direkten Datenaustausch zwischen etablierten Systemen. Diese Anforderungen adressieren Peer-basierte Datenmanagementsysteme (PDMS) in dynamischer und skalierbarer Weise. Anstelle eines globalen Schemas und Schema-Abbildungen zwischen globalem und lokalen Schemata sind Peers untereinander durch Schema-Abbildungen verbunden, über die Anfragen und Daten transformiert und weitergeleitet werden. Solche Abbildungspfade führen allerdings meist zu einem Informationsverlust und vermindern die Qualität der Anfrageergebnisse. Die naive Nutzung sämtlicher vorhandener Abbildungspfade ist ausserdem ineffizient. Wir schlagen für PDMS die Berücksichtigung der Informationsqualität bezüglich Datenquellen, Schema-Abbildungen und Anfrageergebnissen vor und nutzen Konzessionen an die Vollständigkeit zur Verminderung der Antwortzeit. Das Ziel ist ein Optimum zwischen Laufzeit der Anfrage und Qualität der Ergebnisse. Wir illustrieren dies anhand eines konkreten Anwendungsbeispiels.

1 Peer-basierte Datenmanagementsysteme (PDMS)

Der Austausch und die Integration semantisch relevanter Information ist in der heutigen hochdynamischen und komplexen Welt ein drängendes Problem. Dabei liegt die Hauptmotivation der Informationsintegration in einer möglichst umfassenden, also vollständigen Sicht der Welt. Dies erfordert die Einbeziehung möglichst vieler relevanter, aber oft heterogener Datenquellen. Zentralisierte Datenintegrationssysteme (z.B. Data Warehouses) verfolgen die Idee, diesen Anforderungen mit einem globalen, integrierten Schema gerecht zu werden. Der hohe Aufwand zu dessen Bildung und Wartung ist jedoch ein wesentliches Hemmnis für die Skalierbarkeit bezüglich der Anzahl von Quellsystemen.

In der Praxis ist zu beobachten, dass zum Datenaustausch eher ein dezentrales Vorgehen bevorzugt wird. Anfragen sollten im gewohnten Kontext des eigenen Schemas gestellt und über Beziehungen zu ähnlichen benachbarten Systemen bearbeitet werden. Diese Anforderungen adressieren Peer-basierte Datenmanagementsysteme (PDMS). In einem PDMS kann ein Peer sowohl Daten bereitstellen als auch die Rolle eines Mediators einnehmen und Anfragen entgegennehmen. Anfragen werden entsprechend semantischer Beziehungen, sogenannte Mappings, zwischen Peers übersetzt und weitergeleitet (Abbildung 1). Peer-to-Peer-Systeme zum Datenaustausch (P2P), wie z.B. Napster, besitzen im Gegensatz

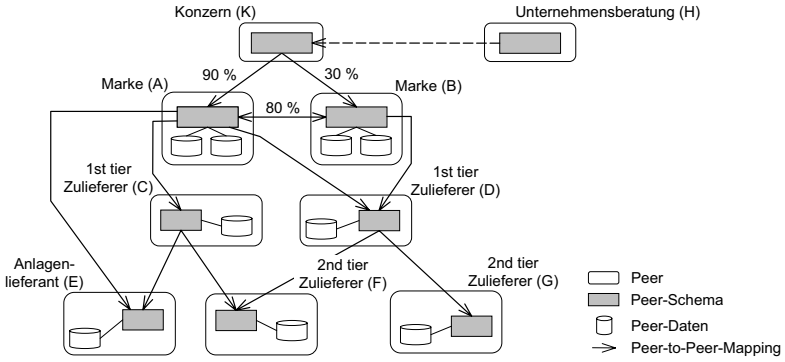

Abbildung 1: Beispielstruktur eines PDMS (Mappings teilweise mit Qualitätsmaß).

zu PDMS einfache Semantik und homogene Schemata und lassen nur einfachste Anfragen zu.

Beispielhafte Anwendungsbereiche für PDMS sind die partnerübergreifende Entwicklung komplexer technischer Produkte, Kooperationen wissenschaftlicher Einrichtungen und ad hoc-Krisenmanagement [HIST03]. Das *Semantic Web* profitiert ebenfalls von PDMS als einer dezentralen Plattform zur Mediation zwischen Ontologien. Durch deren präzise Semantik-Beschreibung vereinfacht sich andererseits die dynamische Erstellung von Mappings für PDMS.

In der Literatur finden sich zwei prinzipiell unterschiedliche Ansätze für PDMS. Das Piazza-Projekt von Halevy et al. [HIST03], die Arbeiten von Bernstein et al. [BGK+02], Aberer et al. [ACMH03] und Lenzerini et al. [CGLR04] propagieren Anfragen und Daten ausschließlich entlang Mappings zwischen den Peers. Dagegen verteilen beim Ansatz mit semantischen Overlay-Netzen sogenannte Super-Peers die Anfragen an passende Peers.

PDMS Piazza: Die Mediation zwischen Schemata eines PDMS ist Thema von [HIST03, HIMT03]. Konzessionen an die Vollständigkeit der Anfrageergebnisse werden zwar angesprochen, aber nicht weiter vertieft. Neue Algorithmen zur Optimierung der Anfrageumformulierung finden sich in [TH04]. Diese sind jedoch unabhängig von der Informationsqualität, deren Berücksichtigung für die Autoren noch offen ist.

Semantic Gossiping: Der Ansatz aus [ACMH03] nutzt Zyklen in Mapping-Netzwerken für die Untersuchung des Informationsverlustes. Die Autoren verwenden ein einfaches Datenmodell und Mappings auf Attributbasis ohne Selektionen. Die Erweiterung dieser Ansätze für ausdrucksstarke Mapping-Formalismen, wie in Piazza beschrieben, ist eine interessante Forschungsperspektive.

Edutella: In diesem sogenannten Schema-basierten P2P-Netzwerk bestehen semantische Overlay-Netze aus Clustern (Super-Peers) semantisch ähnlicher Peers [LNWS03]. Dieser Ansatz verwendet keine direkten Mappings zwischen Peer-Schemata.

Mappings und Informationsverlust: Die Einbeziehung von *Ungenauigkeiten* und *Unsicherheiten* in die Anfragebearbeitung ist eine wichtige Forschungsperspektive [MBDH02].

In [MKIS00] werden ungenaue Mappings zwischen einzelnen Begriffen zugelassen und der dabei auftretende *Informationsverlust* untersucht. Gängige Formalismen für Mappings sind jedoch deutlich ausdrucksstärker. Ebenso erscheint die Anwendung des statistischen Verfahrens aus [AC03] auf Netzwerke solcher Mappings als vielversprechend.

2 Anfragebearbeitung in PDMS

Mappings zwischen Peers sind semantische Verknüpfungen einzelner Schemaelemente und haben den Charakter von Interpretationen [FHP+02] mit *Ungenauigkeiten* und *Unsicherheiten* [MBDH02, AC03]. Transformationen zwischen Datenbeständen sind also mit einem Daten- und Informationsverlust verbunden.

Zur Übersetzung von Anfragen reicht man deren Teilausdrücke entlang der Mappings zwischen den Peers weiter. Agiert ein Peer als Mediator, führt dies eventuell zu einer Verzweigung. Innerhalb eines Zweiges bricht dieser Prozess nach dem Auffinden eines Zyklus oder Peers ohne weitere Mappings ab. Die möglichen Anfragepläne lassen sich aus einem Suchbaum wie in [HIST03] angegeben ermitteln und bestehen aus den Teilausdrücken an den Blättern des Suchbaumes, die gespeicherte Peer-Daten beschreiben. In umfangreichen PDMS kann der Aufbau des komplexen und stark verzweigten Suchbaums beträchtliche Laufzeit erfordern. Ähnlich zu [TH04] ist unser Ziel, diesen Prozess der Anfrageplanung zu verkürzen. Wir schlagen hierfür Konzessionen an das Qualitätskriterium der Vollständigkeit vor und verringern so die Gesamtlaufzeit von Anfragen (Abschnitt 3).

Zusammengefasst ist das Ziel unserer Forschung ein dynamisches Peer-basiertes Datenmanagementsystem (1) mit dezentraler Organisation , (2) bestehend aus einer skalierbaren Anzahl von weitgehend autonomen und gleichberechtigten Systemen (Peers), (3) das Anfragen gegen das Schema *eines* Peers über *alle* relevanten Peers des Netzwerkes verarbeitet und (4) einen Kompromiss zwischen Effizienz und Qualität erlaubt.

3 Informationsqualität und Informationsverlust in PDMS

In diesem Abschnitt erweitern wir die bisherige Sicht auf PDMS um den Aspekt der Informationsqualität (IQ) und beschreiben offene Herausforderungen in diesem Bereich. Informationsqualität ist ein wichtiges Unterscheidungsmerkmal von Datenbeständen und läßt sich als Aggregation der Werte mehrerer IQ-Kriterien auffassen [Na02]. Drei inhaltsbasierte IQ-Kriterien sind in unserem Zusammenhang besonders interessant. *Extensionale Vollständigkeit* (Recall) beschreibt das Verhältnis einer Menge von Objekten zur Anzahl *aller* im PDMS im Zugriff stehenden Objekte. Konkret kann man als Objektmenge die lokalen Daten eines Peers oder das Ergebnis einer Anfrage betrachten. *Intensionale Vollständigkeit* definieren wir als den Anteil der Schemaelemente eines Datenbestandes an den Schemata (Intension) *aller* Peers. *Relevanz* (Precision) ist der Grad an Übereinstimmung eines Anfrageergebnisses mit dem Informationsbedarf eines Nutzers. Im folgenden erläutern wir Problemstellungen und Lösungsansätze in dem bisher beschrie-

benen Kontext PDMS und Informationsqualität.

Informationsqualität und Informationsverlust: Informationsverlust ist ein Sammelbegriff für Einbußen bezüglich der oben beschriebenen IQ-Kriterien [MKIS00]. In [Na02] bilden Konzessionen an die Vollständigkeit der Anfrageergebnisse im Falle von Mediator-basierten Informationssystemen die Grundlage zur Effizienzsteigerung der Anfrageplanung durch Betrachtung der Informationsqualität. PDMS stellen durch den Übergang von der direkten Auswahl von Informationsquellen zu Mapping-Netzen erheblich höhere Anforderungen an die Anfragebearbeitung. Es geht nicht mehr nur um die Auswahl der Informationsquellen, sondern zusätzlich um die Entscheidung über welchen Mapping-Pfad darauf zugegriffen werden soll. Da sich der Informationsverlust entlang solcher Pfade fortpflanzt, ist die Einbeziehung obiger IQ-Kriterien und entsprechender Konzessionen in die Anfragebearbeitung gerade im Falle von PDMS nützlich.

Wir nehmen an, dass Anfragen und Mappings nur aus Selektionen, Projektionen und Joins aufgebaut sind und betrachten zunächst ausschliesslich global-as-view-Mappings. Enthält ein Mapping einen Join, lassen sich die erwartete extensionale und intensionale Vollständigkeit mit Theoremen aus [Na02] berechnen. Selektionen in Mappings vermindern potentiell die extensionale Vollständigkeit, da sie – intuitiv gesehen – die resultierende Informationsmenge einschränken. Projektionen in Mappings können einschränkend auf die intensionale Vollständigkeit wirken. Beträchtliches Potenzial hat eine auf IQ-Kriterien beruhende Beschneidung des Suchbaumes (siehe Beispiel unten). Eine vor der Anfrageumformulierung ausgeführte Mapping-Komposition bietet ebenfalls Gelegenheit, zur Einbeziehung von IQ-Kriterien. Als Ergebnis der IQ-Analyse sind weiterhin das Vorschlagen neuer Peer-Mappings und ein Ranking der Anfrageergebnisse denkbar.

Unsicherheit in Mappings: Unsicherheiten in Mappings beziehen sich auf die Gültigkeit von Ausdrücken, die Elemente von Quell- von Zielschema zueinander in Beziehung setzen. Aufgrund der hohen Ausdrucksmächtigkeit von Mapping-Formalismen ist es eine Herausforderung, Ansätze wie probabilistische Logik anzuwenden [MBDH02].

Beispiel: In das in Abbildung 1 dargestellte PDMS eines Konzerns soll dynamisch das Informations- und Analysesystem H einer Unternehmensberatung integriert werden. Hierzu reicht es aus, ein Mapping zum Konzern-Peer K zu erstellen. Das aggregierte IQ-Maß der über das Mapping $K \rightarrow B$ bereitgestellten Informationen betrage 30 % des IQ-Maßes der für B zugänglichen Informationen. Der Informationsverlust dieses Mappings ist also 70 %. Darum schlägt das PDMS vor, diesen Pfad beim Aufbau des Suchbaums zumindest vorerst zu ignorieren und stattdessen den alternativen Pfad $K \rightarrow A \rightarrow B$ über A zu nutzen. □

4 Zusammenfassung

Wir beschreiben das Problem einer semantisch ausdrucksstarken und qualitätsgesteuerten Anfragebearbeitung in Peer-basierten Systemen zum Datenmanagement (PDMS). Das Ziel ist es, Nutzer *effizient* mit relevanten und qualitativ hochwertigen Informationen zu versorgen, die aus einer hohen Zahl autonomer Quellen stammen.

Vollständigkeit steht als Hauptmotiv der Informationsintegration auch bei PDMS im Mittelpunkt. Wir erläutern, welche Aspekte des PDMS sich auf die Informationsqualität des Anfrageergebnisses auswirken. Als übergreifende Herausforderungen sehen wir die Berücksichtigung des Einflusses von potentiell unvollständigen, inkorrekten und unsicheren Mappings und schlechter Qualität der Daten an den Peers in der Anfragebearbeitung.

Literatur

[AC03] Altareva, E. und Conrad, S.: Statistical analysis as methodological framework for data(base) integration. In: *Proceedings of the International Conference on Conceptual Modeling (ER)*. 2003.

[ACMH03] Aberer, K., Cudré-Mauroux, P., und Hauswirth, M.: The chatty web: Emergent semantics through gossiping. In: *Proceedings of the International World Wide Web Conference (WWW)*. 2003.

[BGK+02] Bernstein, P. A., Giunchiglia, F., Kementsietsidis, A., Mylopoulos, J., Serafini, L., und Zaihrayeu, I.: Data management for peer-to-peer computing: A vision. In: *Proceedings of the ACM SIGMOD Workshop on The Web and Databases (WebDB)*. 2002.

[CGLR04] Calvanese, D., Giacomo, G. D., Lenzerini, M., und Rosati, R.: Logical foundations of peer-to-peer data integration. In: *Proceedings of the Symposium on Principles of Database Systems (PODS)*. 2004.

[FHP+02] Fagin, R., Hernandez, M., Popa, L., Velegrakis, Y., und Miller, R. J.: Translating web data. In: *Proceedings of the International Conference on Very Large Databases (VLDB)*. 2002.

[HIMT03] Halevy, A. Y., Ives, Z., Mork, P., und Tatarinov, I.: Piazza: Data management infrastructure for semantic web applications. In: *Proceedings of the International World Wide Web Conference (WWW)*. 2003.

[HIST03] Halevy, A. Y., Ives, Z., Suciu, D., und Tatarinov, I.: Schema mediation in peer data management systems. In: *Proceedings of the International Conference on Data Engineering (ICDE)*. 2003.

[LNWS03] Löser, A., Nejdl, W., Wolpers, M., und Siberski, W.: Information integration in schema-based peer-to-peer networks. In: *Proceedings of the Conference on Advanced Information Systems Engineering (CAiSE)*. 2003.

[MBDH02] Madhavan, J., Bernstein, P. A., Domingos, P., und Halevy, A. Y.: Representing and reasoning about mappings between domain models. In: *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. 2002.

[MKIS00] Mena, E., Kashyap, V., Illarramendi, A., und Sheth, A. P.: Imprecise answers in distributed environments: Estimation of information loss for multi-ontology based query processing. *Intl. Journal of Cooperative Information Systems*. 9(4):403–425. 2000.

[Na02] Naumann, F.: *Quality-driven query answering for integrated information systems*. Number 2261 in Lecture Notes in Computer Science. Springer. 2002.

[TH04] Tatarinov, I. und Halevy, A.: Efficient query reformulation in peer data management systems. In: *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*. 2004.