Semantische Mediation für heterogene Informationsquellen

Holger Wache

12. September 2003

1 Einleitung

Mit der zunehmenden Verbreitung des Internets gerät die Integration heterogener Informationssysteme immer mehr ins Blickfeld des praktischen, aber auch forschungstechnischen Interesses. Mit der wachsenden Zahl verfügbarer Informationssysteme ergeben sich aber auch neue, zusätzliche Anforderungen an Integrationsansätze, die von den bisherigen nur bedingt erfüllt werden. Zum einen gilt es, die Integration vieler Systeme möglichst einfach zu gestalten. Zum anderen wird eine hohe Skalierbarkeit und Flexibilität des Integrationsansätze gefordert, da im Laufe der Zeit neue Informationssysteme verfügbar werden oder schon eingebundene Systeme sich verändern.

Bisherige, ältere Lösungsansätze (vgl. [1]) konzentrierten sich auf die technischen und syntaktischen Probleme, wobei die heterogenen Semantiken der verschiedenen Informationssysteme vernachlässigt werden. Sie fokussierten also die strukturellen Heterogenitätskonflikte, während die semantischen Heterogenitätskonflikte eine untergeordnete Rolle spielten. Neuere Arbeiten (z.B. [6]) versprechen durch die explizite Berücksichtigung der Semantik eine Vereinfachung der Integration.

In dieser Arbeit wird ein neuartiger Ansatz zur expliziten Repräsentation der Semantik von Informationen vorgestellt, der sich insbesondere durch seine inhärente Skalierbarkeit und Flexibilität von den existierenden Ansätzen abhebt. Daneben gilt es ein Integrationswerkzeug zu entwerfen, welches mit der repräsentierten Semantik adäquat umzugehen weiß. Das Integrationswerkzeug muss nicht nur die strukturellen, sondern insbesondere auch die

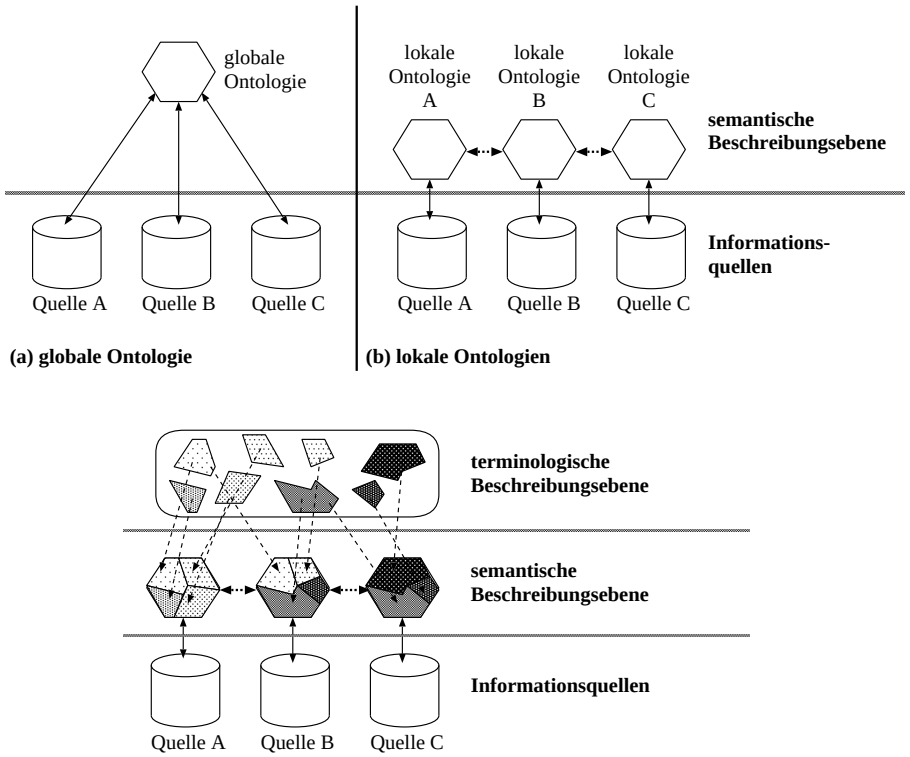

Abbildung 1: Vergleich der Ontologie–basierten Modellierungsansätze

semantischen (Daten-)Heterogenitätskonflikte beseitigen. Hierbei treten insbesondere die Forderungen nach Skalierbarkeit und Flexibilität in den Vordergrund.

2 Die explizite Repräsentation von Semantik der Informationen

Eine wesentliche Voraussetzung für eine Vereinfachung der Integration heterogener Informationssysteme ist sowohl die syntaktische als auch die explizite Repräsentation der Semantik der Informationen. Durch den Vergleich der semantischen und der syntaktischen Beschreibung der Informationen lassen

sich erst alle verschiedenen Heterogenitätskonflikte identifizieren, die bei der Integration auftreten. In jüngster Zeit werden sogenannte Ontologien [3] zu Beschreibung der Semantik eingesetzt. Die Ontologie-basierten Ansätze lassen sich nach [7] dahingehend unterscheiden, ob sich alle Informationssysteme auf eine globale Ontologie beziehen (Single-Ontology-Approach, z.B. [2]), oder jedes Informationssystem durch seine eigene Ontologie charakterisiert wird (Multi-Ontologies-Approach, z.B. [5]). Der häufig verwendete Single-Ontology-Approach hat Defizite bezüglich der Skalierbarkeit und der Flexibilität, während der Multi-Ontologies-Approach die gewünschte Skalierbarkeit und Flexibilität besitzt. Aber die Multi-Ontologies-Approaches weisen dafür erhebliche Probleme beim Vergleich der verschiedenen Ontologien auf, da die verschiedenen Ontologien nach unterschiedlichen Kriterien erstellt wurden. Die Defizite beider Arten von Ansätzen haben sich als gravierende Behinderung herausgestellt, so dass aus praktischer Sicht keiner der beiden Ansätze als adäquat angesehen werden kann.

Um diese Defizite zu beheben, wird ein neuer Ontologie-basierter Ansatz zur semantischen Beschreibung von Informationen vorgestellt. Er kann als eine Mischung der beiden unterschiedlichen Arten Ontologie-basierter Ansätzen verstanden werden. Jedem Informationssystem wird seine eigene Applikationsontologie zugeordnet, wobei aber alle Applikationsontologien auf einem gemeinsamen, globalen Vokabular basieren. Das Vokabular gibt die primitiven Begriffe Informationssystem-übergreifend vor, die dann für die einzelnen Applikationsontologie entsprechend den Konzeptualisierungen ihrer Informationssysteme zu komplexen Begriffen kombiniert werden. Die verschiedenen Applikationsontologien unterstützen die geforderte Flexibilität und Skalierbarkeit in Analogie zu den Ansätzen mit multiplen Ontologien, während das globale Vokabular als Lingua Franca"jedoch die Vergleichbarkeit der Applikationsontologien gewährleistet. Neben diesem Aspekt wird in der Arbeit der neuartige Beschreibungsansatz auf ein wohlfundiertes, logisches Modell gestellt, um eine klare Semantik für die Beschreibungssprache zu erhalten. Das logische Modell weist viele Parallelen zur bekannten Beschreibungslogik auf, weicht aber in einigen entscheidenden Punkten ab.

3 Der regelbasierte Mediator MeCoTA

Neben der syntaktischen und semantischen Beschreibung von Informationen wird für die Integration ein Werkzeug bereitgestellt, dass die Informationen

aus den heterogenen Informationssystemen integriert und kombiniert. Für diesen Zweck bieten sich die Mediator-Wrapper-Architekturen [8] an. Die Wrapper kapseln die Informationssysteme, um eine einheitliche Schnittstelle auf beliebige Informationssysteme zu bieten und die Besonderheiten der Systeme zu vereinheitlichen. Den Mediatoren obliegt dann die Aufgabe, die Informationen zu integrieren, zu kombinieren und die Heterogenitätskonflikte zu beseitigen. Regelbasierte Mediatoren werden durch explizite Integrationsvorschriften in Form von Regeln konfiguriert (vgl. [4]), die festlegen, wie Informationen aus den Informationssystemen zu integrieren und zu kombinieren sind. Aus den bisher in der Literatur diskutierten Formalismen regelbasierter Mediatoren, wie auch allgemein die anderen Integrationsansätze, geht jedoch nicht hervor, welcher Teil für die Beseitigung welches Heterogenitätskonflikts zuständig ist. Die Vermischung erschwert nicht nur die Wartung und Wiederverwendung der Integrationsvorschriften, sondern verletzt insbesondere die Skalierbarkeit und Flexibilität.

In dieser Arbeit wird ein neuer, regelbasierter Integrationsformalismus vorgestellt, der sich durch eine Bipartitionierung der Regelmenge auszeichnet. Hierbei beschreiben Integrationsregeln, wie Informationen aus den verschiedenen Informationssystemen zusammenzuführen sind. Ihre Aufgabe ist die Beseitigung der strukturellen Heterogenitätskonflikte. Die zweite Art von Regeln, die Kontexttransformationsregeln, basieren auf dem Prinzip, dass jede Information in einem Kontext (seines Informationssystems) zu sehen ist. Sie überführen eine Information aus einem Kontext in einen anderen Kontext, wobei die Information gegebenenfalls konvertiert wird. Dabei beseitigen sie die semantischen (Daten-)Heterogenitätskonflikte. Die Bipartitionierung der Regelmenge vereinfacht zunächst die Formulierung der Regeln. Bei einer Integrationsregel brauchen Aspekte der semantisch motivierten Konvertierungen nicht berücksichtigt zu werden; man kann sich ausschließlich auf die eigentliche Integrationsaufgabe, die Integration und Kombination der Informationen, konzentrieren. Mit den Kontexttransformationsregeln werden dann die Informationen vom Kontext der Informationsquelle in den Zielkontext überführt und die notwendigen Konvertierungen veranlasst. Durch die klare Aufgabenverteilung der unterschiedlichen Regelarten vermindert sich der Wartungsaufwand und erhöht sich die Skalierbarkeit und Flexibilität. Darüber hinaus wird die Wiederverwendung erleichtert, da Kontexttransformationsregeln unabhängig von der Anwendungsdomäne formuliert sind und daher zwischen verschiedenen Integrationsszenarien ausgetauscht werden können.

Kontexttransformationsregeln können syntaktisch in einem ähnlichen Formalismus wie die Integrationsregeln formuliert werden. Jedoch weist die Kontexttransformation entscheidende Unterschiede zu den Inferenzen über den Integrationsregeln auf. Augenscheinlich ist die Kontexttransformation mit einer Termersetzung vergleichbar, jedoch gilt für die Kontexttransformation nicht die Symmetrie. Wegen des Fehlens der Symmetrie kann nicht auf die Ergebnisse der Termersetzung zurückgegriffen werden. Deshalb wird ein neues Kalkül für die Kontexttransformation entwickelt, für das die Vollständigkeit und Korrektheit gezeigt wird. Außerdem wird die Kontexttransformation mit der Integration kombiniert. Dazu wird das bekannte Resolutionsprinzip, das die logische Grundlage für die Integration darstellt, modifiziert, indem die Unifikation zur Kontexttransformation verallgemeinert wird. Für die so entstandene CT-Resolution wird ebenfalls die Vollständigkeit und Korrektheit bewiesen. Die CT-Resolution stellt die formale Basis für den in dieser Arbeit entwickelten Mediator dar.

4 Kontakt

Holger Wache
Technologie-Zentrum Informatik
Universität Bremen
Postfach 330440
28334 Bremen
Tel: +49 (0)421 218-7838
Fax: +49 (0)421 218-7196
e-mail: wache@tzi.de
http://www.tzi.de/∼wache

Literatur

[1] Stefan Conrad. *Föderation heterogener Datenbanken.* Datenbank-Management. Interest-Verlag, 1998.

[2] D. Fensel, J. Angele, S. Decker, M. Erdmann, H.-P. Schnurr, S. Staab, R. Studer, and A. Witt. On2broker: Semantic-based access to information sources at the www. In P. De Bra and J. J. Leggett, editors, *Proceedings of the World Conference on the WWW and Internet (WebNet)*, pages

366–371, Charlottesville, VA, USA, 25-30 Oktober 1999. Association for the Advancement of Computing in Eduction (AACE).

[3] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[4] Alon Y. Levy. Kap.: Logic-based techniques in data integration. In Jack Minker, editor, *Logic Based Artificial Intelligence*, pages 575–597. Kluwer Academic Publishers, 2000.

[5] E. Mena, A. Illarramendi, V. Kashyap, and A. P. Sheth. Observer: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distributed and Parallel Databases*, 8(2):223–271, 2000.

[6] C. F. Naiman and A. M. Ouksel. A classification of semantic conflicts in heterogeneous database systems. *Journal of Organizational Computing*, pages 167–193, 1995.

[7] Holger Wache, Thomas Vögele, Ubbo Visser, Heiner Stuckenschmidt, Gerhard Schuster, Holger Neumann, and Sebastian Hübner. Ontology-based integration of information — a survey of existing approaches. In A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, and M. Uschold, editors, *Proceedings of the IJCAI-Workshop Ontologies and Information Sharing*, pages 108–117, Seattle, WA, September 2001.

[8] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, March 1992.