Exposé zur Diplomarbeit

Anfrageübersetzungen in SPARQL für verteilte Quellen

von Alexander Musidlowski

Betreuung: Ulf Leser, Bastian Quilitz

1 Motivation

Das Internet bietet heutzutage viele unstrukturierte Informationen aus verschiedensten Quellen an. Das Semantic Web hat sich zum Ziel gesetzt, die Informationen aus den unterschiedlichen Quellen zu integrieren und zu kombinieren [5]. Dazu müssen die Informationen auch automatisiert verarbeitet werden können. Die Voraussetzung dafür ist, dass eine klare Semantik der Daten vorliegt.

Deshalb müssen die Unterschiede in der Bedeutung von Begriffen in verschiedenen Quellen (Heterogenität der Daten) überwunden werden. In der Version des Semantic Web geschieht durch die Verwendung einheitlicher Vokabulare beziehungsweise Ontologien in den Quellen. Mit Hilfe von Ontologien versucht man das Wissen zu strukturieren und so auf eine einheitliche formale Darstellung zu bringen. Nach Thomas Gruber ist eine Ontologie „an explicit specification of a conceptualization“[4]. Diese formalen Spezifikationen sind zumeist auf einen Wissensbereich beschränkt. Ein Beispiel dafür ist FOAF - Friend of a Friend.

> The Friend of a Friend (FOAF) project is creating a Web of machine-readable pages describing people, the links between them and the things they create and do. [1]

Das **R**esource **D**escription **F**ramework (RDF [6]) wird genutzt um Daten strukturiert darzustellen. RDF ist ein graphbasiertes Datenmodell zur Repräsentation von Daten in

Form von Aussagen mit Subjekt, Prädikat und Objekt. Eine Menge von Aussagen bildet eine formale Datenbasis, welche die Grundlage im Semantic Web bilden. Deklarative Anfragen gegen diese Daten können mit SPARQL (**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage [9]) gestellt werden. Die Struktur einer RDF Datenbasis wird als Vokabular beziehungsweise Ontologie bezeichnet und wird technisch u.a. über RDF Schema [3] realisiert.

Im vorliegenden Szenario gibt es verteilte, autonome Quellen, wodurch verschiedene Vokabulare respektive Ontologien entstehen, die zu einer Heterogenität der Daten führen. In [7] steht: „Generell ist zu beobachten, dass der Grad der Heterogenität zwischen Datenquellen mit dem Grad der Autonomie der Quellen zunimmt."

Für verteilte Quellen kann man die Funktionen des Programms DARQ verwenden, welches SPARQL Anfragen entgegennimmt und diese mit Hilfe bekannter Quellen gleichen Vokabulars wie in der Anfrage, beantwortet.

2 Problemstellung

Das Ziel dieser Arbeit ist die Integration verteilter, autonomer Quellen bei einer SPARQL Anfrage. Eine Lösung des Problems ist der integrierte Zugriff auf die RDF Datenquellen mit einer SPARQL Anfrage. Das Programm DARQ ist in der Lage Anfragen in SPARQL aufzusplitten und auf die entsprechenden Quellen zu verteilen. DARQ setzt voraus, dass die verwendeten Vokabulare in der Anfrage und den Quellen übereinstimmen. Deshalb besteht die Aufgabe darin, DARQ so zu erweitern, dass eine Anfrage beantwortetet werden kann, obwohl die Quellen unterschiedliche RDF Schemata verwenden.

Die Herausforderung ist, aus denen von DARQ bereits aufgesplitteten Anfragen neue äquivalente Anfragen in alternativen Vokabularen zu generieren.

Dabei ergeben sich Teilprobleme wie die, ein geeignetes, universelles Format für die Korrespondenzen (mapping) zwischen den Vokabularen zu finden und zu implementieren. Dieses Format soll durch den Nutzer zu erstellen sein und dennoch möglichst viele Konfliktarten abbilden können.

Diese Konflikte lassen sich auf die Designautononmie der Quellen zurückführen, welche Ursache für strukturelle, semantische und schematische Heterogenität in den Schemata ist. Nach [7] zählen diese Arten der Heterogenität zu den schwierigsten in der Informationsintegration. In [11] werden die Probleme und Ansätze zur Überwindung der Heterogenität (Mapping) ausführlich erläutert. Das Finden von Mappings (auch Schema Matching genannt) ist nicht die Aufgabe dieser Diplomarbeit. Stattdessen gehen wir

davon aus, dass die Korrespondenzen zwischen den Vokabularen der Quellen von einem Nutzer definiert und dem Programm zur Verfügung gestellt werden.
Nach der Klassifiktion von Wache (vgl. [12]) gibt es u.a. die folgende Konflikte:

- Bezeichnerkonflikte
- Subsumptionskonflikte
- Konflikte mit multilateralen Attributkorrespondenzen
- Einheiten- und Skalierungskonflikte.

Diese Konflikte sollen durch die Erweiterung von DARQ gelöst werden. Dazu werden die Korrespondenzen bei einer Anfrage ausgewertet, um festzustellen, ob weitere Quellen mit einem anderen Vokabular ebenfalls Daten beinhalten, die zur Beantwortung der Anfrage sinnvoll erscheinen.

DARQ besitzt bereits einen Algorithmus zur Optimierung von Anfrageplänen, so dass von einer Datenquelle nur die notwendigen Daten abgefragt werden. Dieser Algorithmus hilft jedoch nicht weiter, wenn eine Anfrage mehrmals nacheinander gestellt wird. Daher soll DARQ um einen Cache erweitert werden, in dem die Antworten der Datenquellen zwischengespeichert werden, wodurch die Antwortzeit als auch die Netzwerklast verringert werden. Die im Cache gespeicherten Resultate können wiederum als Datenquelle für neue Anfragen dienen (query containment).

3 Vorgehen/Lösungsansatz

Die vorhandenen Mappings werden über eine Datei eingelesen, die der Nutzer vorher erstellt hat. In dieser Datei stehen u.a. Regeln in der Form Regelkopf:Regelkörper. Im Regelkörper steht die Bedingung und im Regelkopf die Konsequenz, wie im Beispiel mit den Prädikaten Preis_in_Euro und Preis_in_Dollar zu sehen ist.

Preis_in_Euro(x,z): multiply(z, Preis_in_Dollar(x,y), 1.50)

Das Verfahren basiert auf dem Prinzip von Global-as-View sofern man die Anfrage als globales Schema und die Schemata der Quellen als lokal ansieht. Beispielsweise wird ein Subjekt des RDF Tripels aus der Anfrage als Regelkopf gesucht und in einer neuen Anfrage durch die Regelkonsequenz ausgetauscht. Aus den vorhandenen Mappings bzw. den Regelkonsequenzen und der Anfrage des Nutzers werden neue Anfragen für die einzelnen Quellen erzeugt. Der Ergebnisse der neuen Anfragen werden anschließend zusamengeführt.

Bei den o.g. Bezeichnerkonflikten genügt der Austausch des entsprechenden Bezeichners durch sein Synonym. Mit der Selektion einer Teilmenge werden Subsumptionskonflikte gelöst, sofern sich diese Menge selektieren lässt. Wenn die Daten in unterschiedlich vielen Attributen vorliegen, spricht man von Konflikten mit multilateralen Attributkorrespondenzen. Deshalb muss eine Konkatenierung bzw. eine Zerlegung der Zeichenketten stattfinden. Während Erstere relativ problemlos zu implementieren ist, bereitet die Zerlegung Schwierigkeiten, da diese nach vorher festgelegten Vorschriften geschehen muss. Zum Lösen von Einheiten- und Skalierungskonflikte muss beim Mapping der Umrechnungsfaktor angegeben sein, so dass die Werte vergleichbar sind. Es existieren weitere Konflikte, bei denen noch zu prüfen ist, ob diese sich lösen lassen. Dazu gehören die Datentypkonflikte. Hierbei stellt sich das Problem, ob die Datentypen im Mapping angegeben werden können. Alternativ könnte man auch eine generelle Überprüfung der Datentypen implementieren, wobei sich hier die Frage nach der Effizienz stellt. Ein weiterer Konflikttyp sind die Repräsentationskonflikte, welche das Analogon zu Einheiten- und Skalierungskonflikten von Zeichenketten darstellen. Die Fragestellung hierbei ist, ob sich eine bijektive Abbildung als Funktion oder Tabelle im Mapping umsetzen lässt.

Da bei der Integration verschiedener Quellen nicht sichergestellt werden kann, dass in jeder Quelle alle Datensätze vollständig vorhanden sind, gilt die *Open World Assumption*.

Es gibt andere Ansätze, die in Projekten wie Observer, SIMS [2] oder PICSEL [10] umgesetzt werden. Das Observerprojekt setzt beispielsweise voraus, dass die Ontologien der einzubindenden Quellen bereits bekannt sind. Diese werden dann in einer Beschreibungslogik nachgebaut. Anschließend wird mit einer erweiterten relationalen Algebra das Mapping zwischen den nachgebauten Ontologien und den Quellontologien erstellt [8]. Der Nutzer kann bei einer Anfrage zwischen den Ontologien wählen.

Literatur

[1] *The friend of a friend project*. http://www.foaf-project.org/

[2] Arens, Y. ; Knoblock, C. : SIMS: Retrieving and integrating information from multiple sources. In: *SIGMOD Rec.* 22 (1993), Nr. 2, S. 562–563. http://dx.doi.org/http://doi.acm.org/10.1145/170036.171566. – DOI http://doi.acm.org/10.1145/170036.171566. – ISSN 0163–5808

[3] Brickley, D. ; Guha, R. : *RDF Schema*. Recommendation. World Wide Web Consortium (W3C), Februar 2004. http://www.w3.org/TR/rdf-schema/

[4] Gruber, T. R.: A translation approach to portable ontology specifications. In: *Knowledge Acquisition* 5 (1993), Nr. 2, 199–220. http://tomgruber.org/writing/ontolingua-kaj-1993.htm. – ISSN 1042–8143

[5] Herman, I. : *W3C Semantic Web Activity*. http://www.w3.org/2001/sw/. Version: 01 2008

[6] Klyne, G. ; Carroll, J. J.: *Resource Description Framework (RDF): Concepts and Abstract Syntax*. W3C Recommendation. World Wide Web Consortium (W3C), Februar 2004. http://www.w3.org/TR/rdf-concepts/

[7] Leser, U. ; Naumann, F. : *Informationsintegration - Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. 1. Auflage. dpunkt.verlag, 2007. – ISBN 3898644006

[8] Mena, E. ; Illarramendi, A. ; Kashyap, V. ; Sheth, A. : OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. In: *International journal on Distributed And Parallel Databases (DAPD)* 8 (2000), April, Nr. 2, S. 223–272

[9] Prud'hommeaux, E. ; Seaborne, A. : *SPARQL Query Language for RDF*. Recommendation. World Wide Web Consortium (W3C), Januar 2008. http://www.w3.org/TR/rdf-sparql-query/

[10] Rousset, M.-C. ; Reynaud, C. : PICSEL and Xyleme: Two Illustrative Information Integration Agents. In: Klusch, M. (Hrsg.) ; Bergamaschi, S. (Hrsg.) ; Edwards, P. (Hrsg.) ; Petta, P. (Hrsg.): *AgentLink* Bd. 2586, Springer (Lecture Notes in Computer Science). – ISBN 3–540–00759–8, 50-78

[11] Stuckenschmidt, H. ; Harmelen, F. van: *Information Sharing on the Semantic Web*. Springer Verlag, 2004. – ISBN 3540205942

[12] Wache, H. : *Semantische Mediation für heterogene Informationsquellen*, Technologie-Zentrum Informatik Universität Bremen, Diss., September 2003. http://www.cs.vu.nl/~holger/