Exposé zur Diplomarbeit

Verwendung von Graphindexen zur Optimierung von SPARQL-Anfragen

Christian Rothe

Betreuung: Ralf Heese, Prof. Ulf Leser
HU Berlin, Institut für Informatik

1 Hintergrund

Semantic Web beschäftigt sich damit, Informationen für Maschinen lesbar und damit auch auswertbar zur Verfügung zu stellen. Das *Resource Description Framework (RDF)* [MM04] ist ein Ansatz, dieses Ziel zu realisieren. Informationen werden dazu in kleine Einheiten zerlegt und als Tripel von Subjekt, Prädikat und Objekt erfasst. Das Subjekt bezeichnet im Tripel die zu beschreibende Sache (die *Ressource*), das Prädikat erfasst eine Eigenschaft des Subjekts und das Objekt gibt schließlich den Wert an, den die Ressource bezüglich der Eigenschaft annimmt. Dieser Wert kann ein *Literal* oder wieder eine Ressource sein. Ressourcen sind im *RDF eindeutig* mit einem *Uniform Resource Identifier (URI)* bzw. *Internationalized Resource Identifier (IRI)* gekennzeichnet, während Literale beliebige Zeichenketten sind. So entspricht die Aussage ,,Das Buch hat die Farbe grün." etwa dem Tripel (Buch, Farbe, grün). Ressourcen können mehrere Eigenschaften oder auch Eigenschaften mit mehreren Werten (z.B. Autor) haben. Es ist auch möglich, Aussagen über Tripel zu treffen (*Reification*), so dass ein Tripel an die Stelle des Objektes in einem anderen Tripel treten kann. Durch die Zusammenfassung mehrerer Tripel erhält man eine *RDF-Datenmenge*, die auf natürliche Weise als gerichteter Multigraph mit Bezeichnern an Knoten und Kanten dargestellt werden kann.
Es gibt verschiedene Anfragesprachen, mit denen Informationen aus den RDF-Daten gewonnen werden können. Eine dieser Sprachen, SPARQL [PS06], bietet die Möglichkeit, Schablonen von Subgraphen zu erstellen, die von einem Compiler im Datengraph gesucht werden. Dabei können sowohl per URI eindeutig identifizierte Knoten und Kanten als auch variabel gehaltene Ressourcen Teil des Suchmusters sein. Aus den gefundenen Subgraphen können dann Informationen ausgelesen werden. Ebenso ist die Erzeugung neuer Gra-

phen möglich.

2 Problem

Jeder Suche nach Vorkommen eines Subgraphmusters im Datengraph liegt der Vergleich zwischen Anfragegraph und (theoretisch) allen Subgraphen des Datengraphs zugrunde – also das Subgraphisomorphieproblem. Die Lösung des Problems ist sehr aufwändig und rechenintensiv (vgl. [Car02]), daher sucht man nach Möglichkeiten das Problem zu vereinfachen. Ein Lösungsansatz ist die Verwendung von Indexen, also die Erfassung verschiedener, in irgendeinem Sinne gleichartiger Subgraphen des Datengraph. Die Literatur kennt mehrere verschiedene Indexierungsstrategien, exemplarisch seien der $D(k)$-Index von Qun, Lim und Ong [QLO03], der Fast Index von Cooper et. al. [CSF+01] oder der Ansatz von Stuckenschmidt et. al. [SVHB04] für verteilte RDF Repositories. Ein tiefer gehender Vergleich verschiedener Strategien wird Teil der Arbeit sein. Das Hauptaugenmerk gilt jedoch materialisierten SPARQL-Anfragen, d.h. vorab erfassten Anfragemustern und ihren Vorkommen in den Daten. Ziel ist durch die Verwendung von Indexen und der darin enthaltenen Informationen, die Zahl der Isomorphietests und damit die Gesamtbearbeitungszeit der Anfrage zu reduzieren.

3 Lösungsansatz

In meiner Studienarbeit [Rot06] wurde das Problem formuliert und theoretisch untersucht. Ausserdem habe ich die Konzepte Abhängigkeit und Selektivität für Graphindexe eingeführt sowie erste Lösungsansätze entwickelt und implementiert. Ausgehend von diesen Ergebnissen sollen die Eigenschaften von Indexen und ihre Verwendungsmöglichkeiten für die Anfragebearbeitung weiter untersucht werden. Dabei soll das Konzept der Abhängigkeit nicht mehr nur auf Daten- sondern auch auf Musterebene betrachtet werden. Damit einhergehend wird es einen neuen Ansatz für die Berechnung der Selektivität geben. Neben der Selektivität sollen weitere statistische Merkmale wie Bezeichnerhäufigkeiten bestimmt und auf ihre Verwendbarkeit hin untersucht werden.
Kernbestandteile bleiben die Zusammenfassung von Indexen zu Überdeckungen und die Auswahl optimaler Überdeckungen gemäß einer Kostenfunktion, welche sich unter anderem auf die geschätzte Selektivität und die Größe der Überdeckung stützen wird. Dabei werden nur solche Indexe betrachtet werden, deren Muster vollständig im Anfragemuster zu finden ist.

Zur Evaluierung der theoretischen Erkenntnisse soll ein flexibler Rahmen für die Anfragebearbeitung entwickelt werden, der den Vergleich zwischen indexbasierter und indexfreier Anfrageausführung ermöglicht. Wie in der Studienarbeit soll dabei die Jena-Bibliothek [CDD+04] zum Einsatz kommen. Die Entwicklung dieses Rahmenkonstruktes wird Grundlage einer umfassenderen Implementation der betrachteten Algorithmen sein.

Literatur

[Car02] Jeremy J. Carroll. Matching rdf graphs. In Ian Horrocks and James A. Hendler, editors, *Proceedings of the First International Semantic Web Conference*, volume 2342 of *Lecture Notes in Computer Science*, pages 5–15. Springer, June 2002.

[CDD+04] Jeremy J. Carroll, Ian Dickinson, Chris Dollin, Dave Reynolds, Andy Seaborne, and Kevin Wilkinson. Jena: implementing the semantic web recommendations. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *WWW (Alternate Track Papers & Posters)*, pages 74–83. ACM, 2004.

[CSF+01] Brian F. Cooper, Neal Sample, Michael J. Franklin, Gísli R. Hjaltason, and Moshe Shadmon. A fast index for semistructured data. In Peter M. G. Apers, Paolo Atzeni, Stefano Ceri, Stefano Paraboschi, Kotagiri Ramamohanarao, and Richard T. Snodgrass, editors, *VLDB*, pages 341–350. Morgan Kaufmann, 2001.

[MM04] Frank Manola and Eric Miller. Rdf primer, February 2004. W3C Recommendation.

[PS06] Eric Prud'hommeaux and Andy Seaborne. Sparql query language for RDF, April 2006. W3C Working Draft.

[QLO03] Chen Qun, Andrew Lim, and Kian Win Ong. D(k)-index: An adaptive structural summary for graph-structured data. In Alon Y. Halevy, Zachary G. Ives, and AnHai Doan, editors, *SIGMOD Conference*, pages 134–144. ACM, 2003.

[Rot06] Christian Rothe. Auswahl von Graphindexen zur Optimierung von SPARQL-Anfragen. Studienarbeit, Juli 2006.

[SVHB04] Heiner Stuckenschmidt, Richard Vdovjak, Geert-Jan Houben, and Jeen Broekstra. Index structures and algorithms for querying

distributed rdf repositories. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *WWW*, pages 631–639. ACM, 2004.