Exposé zur Studienarbeit

Ausnutzung von Graphindexen zur Optimierung von SparQL-Anfragen

Christian Rothe

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

1 Hintergrund

Im Rahmen des Semantic Web und bei Verwendung des RDF [MM04] entstehen Graphen, die verschiedene Ressourcen, deren Eigenschaften und deren Beziehungen untereinander darstellen. Dabei können praktisch beliebig große Netzwerke in Form gerichteter Multigraphen, in denen sowohl Knoten als auch Kanten Label erhalten können – welche bei allen inneren Knoten und einigen Blättern in Form von URIrefs eindeutig ist.
Mit Hilfe von Anfragesprachen wie SparQL [PS06] können diese Graphen nun nach Teilgraphen durchsucht werden. Dabei können sowohl eindeutig identifizierte Knoten und Kanten Teil des Suchgraphs sein als auch – sinnvollerweise – variabel gehaltene Teile. So können einzelne Informationen oder ganze Subgraphen aus einem oder auch mehreren Datengraphen extrahiert werden. Ebenso kann man einen neuen Graphen anhand der Informationen aus den durchsuchten Graphen erzeugen.
Analog zu solchen Suchanfragen können mit vergleichbaren Suchmustern auch Indizes erstellt werden, welche die Vorkommen einzelner – idealerweise häufiger – Muster speichern und dem schnellen Zugriff verfügbar machen.

2 Problem

Das zugrunde liegende Problem bei der Suche mit SparQL in einem Graphen ist der Isomorphietest zwischen dem Anfragegraph und (theoretisch) allen Subgraphen des Datengraphen. Ein solcher Test ist sehr aufwändig und rechenintensiv (vgl. [Car02]), daher soll die Zahl der zu testenden Subgraphen des Datengraphs durch Verwendung eines Indexes, der alle Vorkommen einer gegebenen Menge von Mustern enthält, reduziert werden, damit der

teure Isomorphietest weniger oft bzw. auf einem kleineren Teil der Daten angewendet werden muss.

3 Lösungsansatz

Mit Hilfe von Indexstrukturen soll der Suchraum für die Isomorphietests verringert werden. Im Idealfall ist eine Suchanfrage deckungsgleich mit einem der angelegten Indexmuster und alle passenden Subgraphen können sofort mit Hilfe eben dieses Indexes ausgegeben werden. Bei weitem häufiger jedoch wird die Situation auftreten, dass ein Teil eines Indexmusters mit einem Teil des Anfragemusters übereinstimmt. Hat man nun bei Berücksichtigung aller Indizes mehrere solcher Übereinstimmungen, so kann eine teilweise oder sogar vollständige Überdeckung des Anfragemusters durch die Indexmuster erfolgen. Dadurch soll die Zahl potentieller Kandidaten für ein Vorkommen des Suchmusters im Datengraph reduziert werden.
Ziel der Studienarbeit soll sein, verschiedene mögliche Index- und Anfragemuster zu untersuchen, dabei verschiedene Überdeckungen zu bestimmen und zu analysieren, und schließlich Strategien für (sub-)optimale Überdeckungen zu konstruieren. Dabei soll anhand eines noch genauer zu definierenden Begriffs der Selektivität eines solchen Indexes optimiert werden. Auch eine formale Definition des Problems soll Teil der Arbeit sein.
Abschließend soll eine einfache Implementation des entwickelten Algorithmus erfolgen.

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.

[HD05] Andreas Harth and Stefan Decker. Optimized index structures for querying rdf from the web. In *LA-WEB*, pages 71–80. IEEE Computer Society, 2005.

[MAYU03] Akiyoshi Matono, Toshiyuki Amagasa, Masatoshi Yoshikawa, and Shunsuke Uemura. An Indexing Scheme for RDF and RDF Schema based on Suffix Arrays. pages 151–168, 2003.

[MM04] Frank Manola and Eric Miller. RDF Primer, February 2004. W3C Recommendation.

[PS06] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF, February 2006. W3C Working Draft.

[YYH04] Xifeng Yan, Philip S. Yu, and Jiawei Han. Graph indexing: A frequent structure-based approach. In Gerhard Weikum, Arnd Christian König, and Stefan Deßloch, editors, *SIGMOD Conference*, pages 335–346. ACM, 2004.