Samenvatting

Het semantische web-datamodel, RDF, wordt in toenemende gebruikt in verschillende domeinen zoals de life sciences en publishing, en is uitgegroeid tot standaard voor wereldwijde gegevensstandaardisatie en interoperabiliteit. RDF biedt flexibiliteit voor gebruikers om gegevens weer te geven en te ontwikkelen zonder dat daar een schema voor nodig is, zodat de wereldwijde RDF-graaf (het "semantische web") door iedereen kan worden uitgebreid op eigen initiatief. Deze flexibiliteit brengt een aantal problemen met zich mee in systemen die grote hoeveelheden RDF-gegevens beheren, omdat het de behoefte aan een schema en het begrip van structuur in de RDF-gegevens minder benadrukt. In de eerste plaats verhoogt het ontbreken van schema-informatie de complexiteit van query-optimalisatie, zodat in de praktijk RDF database-systemen een veel kleiner gedeelte van de zoekruimte kunnen bekijken, en er slechtere en dus veel langzamere query-plannen gevonden worden. Daarnaast zorgt de lage gegevenslokaliteit ervoor dat het gebruik van geavanceerde fysieke opslagoptimalisaties voor relationele databases, zoals geclusterde indexering en gegevenspartitionering, niet mogelijk is. Tot slot is het door een gebrek aan schema-inzicht moeilijk voor eindgebruikers om goede SPARQL queries te schrijven. Dit proefschrift gaat in op elk van deze drie problemen. We ontdekken en exploiteren het feit dat echte RDF datasets in vrij hoge mate tabulair gestructureerd zijn. Het automatisch herkennen van zulke structuur maakt het mogelijk RDF-opslag efficiënter en gebruiksvriendelijker te maken.

Een belangrijke constatering van dit proefschrift is dat aan het begrip "schema" een verschillende betekenis toegekend wordt in het semantisch web dan in databases. Binnen het semantisch web verwijst "schema" naar ontologieën en vocabulaires die worden gebruikt om concepten op een generieke manier te beschrijven, zodat die concepten in vele situaties en toepassingen (her-)bruikbaar zijn. In databases verwijst "schema" naar iets heel anders, namelijk naar de specifieke structuur van gegevens in een enkele dataset. Wij betogen dat beide betekenissen van een schema waardevol zijn. Semantische schema's zouden een waardevolle toevoeging kunnen zijn aan relationele databases: de semantiek van een tabel (de entiteit die het kan vertegenwoordigen) en van zijn kolommen en relaties wordt expliciet gemaakt. Dit kan de integratie van gegevens uit verschillende databases vergemakkelijken. Relationele schema's zijn ook waardevol voor semantische webgegevens: de opslag van RDF-gegevens op een schijf of in geheugen kan er beter mee georganiseerd worden zodat RDF-databases betere optimalisaties kunnen uitvoeren, en gebruikers kunnen beter begrijpen welke attributen werkelijk in een RDF-dataset aanwezig zijn.

Dit proefschrift stelt nieuwe technieken voor om automatisch een zogenaamd

"emergent" relationeel schema af te leiden van een RDF-dataset. Het resultaat is een compact en nauwkeurig relationeel schema waarin tabellen, kolommen en relaties korte namen krijgen die makkelijk voor mensen leesbaar zijn. Dit emergente relationele schema is niet alleen nuttig om mensen de structuur RDF-gegevens beter te laten begrijpen; het kan ook de computer helpen om een RDF database-systeem efficiënter te maken. In concreto, het gebruik van een emergent, relationeel schema maakt het mogelijk om RDF-opslag compacter en sneller toegankelijk te maken. Daarnaast helpt het bij het verminderen van het aantal joins (met name zelf-joins) dat nodig is voor SPARQL-queries en het verlagen van de complexiteit van de query-optimalisatie. Dit leidt tot een significante prestatieverbetering in RDF-systemen. Onze methode biedt een veelbelovend perspectief op het ontwikkelen van een efficiënte RDF-opslag die zich kan meten met relationele systemen qua prestatie zonder in te leveren op de flexibiliteit die het RDF-model biedt.

Naast de bijdragen aan het ontwikkelen van hoogwaardige RDF-opslag die gebruik maakt van het automatisch afgeleide, emergente, relationele schema geven we in dit proefschrift ook inzichten en methodes voor het evalueren van de prestaties van RDF-systemen. We hebben een schaalbare datagenerator ontwikkeld die synthetische RDF-graph gegevens kan genereren met scheve datadistributies en plausibele structurele correlaties. Deze gegevensgenerator kan dankzij parallellisatie via Hadoop / MapReduce, een sociale netwerkstructuur genereren met miljarden gebruikersprofielen, verrijkt met interesses, labels, berichten en opmerkingen met behulp van een cluster van alledaagse hardware. De gegenereerde data vertonen ook interessante, realistische waardecorrelaties (bijv. namen vs. landen), structurele correlaties (bijv. vriendschappen versus locatie) en statistische verdelingen ("power laws") die vergelijkbaar zijn met een echt sociaal netwerk zoals Facebook. Deze gegevensgenerator vormt nu de kern van een industriële benchmark, de LDBC Social Network Benchmark (SNB), die is ontworpen om RDF-graph systemen te evalueren.