Abstract

The main semantic web data model, RDF, has been gaining significant traction in various domains such as the life sciences and publishing, and has become the unrivaled standard behind the vision of global data standardization and interoperability over the web. This data model provides the necessary flexibility for users to represent and evolve data without prior need of a schema, so that the global RDF graph (the semantic web) can be extended by everyone in a grass-roots and pay-as-you-go way. However, as identified in this thesis, this flexibility which de-emphasizes the need for a schema and the notion of structure in the RDF data poses a number of data management issues in systems that manage large amounts of RDF data. Specifically, it leads to (i) query plans with excessive join complexity which are difficult to optimize, (ii) low data locality which blocks the use of advanced relational physical storage optimizations such as clustered indexing, data partitioning, and (iii) a lack of schema insight which makes it harder for end-users to write SPARQL queries with non-empty-results.

This thesis addresses all three problems. We uncover and exploit the fact that real RDF data, while not as regularly structured as relational data, still has the great majority of triples conforming to regular patterns. Recognizing this structure information allows RDF stores to become both more efficient *and* easier to use. An important take-away from this thesis is that the notion of “schema” is understood differently in semantic web than in databases. In semantic web “schema” refers to ontologies and vocabularies which are used to describe entities in terms of their properties and relationships in a generic manner, that is valuable across many different application contexts and datasets. In databases, “schema” means the properties of data stored in a single database. We argue both different notions of schema are valuable. Semantic schemas could be a valuable addition to relational databases, such that the semantics of a table (the entity it may represent) and of its columns and relationships is made explicit. This can facilitate data integration. Relational schemas are valuable for semantic web data, such that RDF stores can better organize data on disk and in memory, SPARQL engines can do better optimizations, and SPARQL users can better understand the nature of an RDF dataset. This thesis concentrates on these latter points. Concretely, we propose novel techniques to automatically derive a so-called emergent relational schema from an RDF dataset that recovers a compact and precise relational schema with high triple coverage and short human-readable labels. Beyond the use of the derived emergent relational schema for conveying the structure information of RDF dataset to users and allowing humans to understand RDF dataset better, we have exploited this emergent

schema internally inside the RDF system (in storage, optimization, and execution) to make RDF stores more efficient. In particular, using emergent relational schema allows to make RDF storages more compact and faster-to-access, and helps reducing the number of joins (i.e., self-joins) needed in SPARQL query execution as well as the complexity of query optimization, showing significant performance improvement in RDF systems. This approach opens a promising direction in developing efficient RDF stores which can bring RDF-based systems on par with relational-based systems in terms of performance without losing any of the flexibility offered by the RDF model.

Besides the contributions on developing high performance RDF stores using the automatically derived emergent relational schema, in this thesis, we also provided insights and materials for evaluating the performance and technical challenges of RDF/graph systems. Particularly, we developed a scalable graph data generator which can generate synthetic RDF/graph data having skewed data distributions and plausible structural correlations of a real social network. This data generator, by leveraging parallelism though the Hadoop/MapReduce paradigm, can generate a social network structure with billions of user profiles, enriched with interests/tags, posts, and comments using a cluster of commodity hardwares. The generated data also exhibited interesting realistic value correlations (e.g., names vs countries), structural correlations (e.g., friendships vs location), and statistical distributions (e.g., power-law distribution) akin to a real social network such as Facebook. Furthermore, the data generator has been extended and become a core ingredient of an RDF/graph benchmark, LDBC Social Network Benchmark (SNB), which is designed to evaluate technical challenges and solutions in RDF/graph systems.