Transparent Querying of Distributed Linked Data

Martin Svoboda, Irena Mlýnková
{svoboda,mlynkova}@ksi.mff.cuni.cz

Charles University in Prague
XML and Web Engineering Research Group

Introduction

… what is this poster presentation about?

The concept of **Linked Data** appeared recently in order to allow publishing data on the Web in a way more suitable for automated processing by programs and not only traditional users. Despite the research effort in recent years, several questions in the area of **indexing and querying** remain open, not only since the amount of Linked Data globally available significantly increases each year.

Our ongoing research effort should result in a **proposal of a new querying system** dealing with disadvantages of the existing approaches. This poster should reveal the **model and architecture** of this system and discuss several related issues and aspects we are dealing with.

Objectives

… what is the goal of our ongoing research effort?

The querying framework we are attempting to propose should focus especially on four main problems of the existing approaches we identified in our previous work — they are **data distribution, scaling, dynamicity and quality**. It would be out of our possibilities to fully cope with all of them, but we still need to consider all of them and at least attempt to do it.

This framework should enable **distributed querying**, which means that it allows working with locally stored data as well as distributed data on remote sources. The query evaluation process itself is also distributed.

One of the most important ideas of this framework is that we do not specify data we want to work with during the time of querying, but in advance and only once when a **distributed database** as a set of such data is designed.

This allows **transparent querying** for users of such databases. This assumption also means that we can build and adjust these databases with respect to the **knowledge of data**. Thus, we are able to select appropriate storages or indexing structures in order to get more efficient query processing.

Challenges

… what are the open problems of Linked Data querying?

Distribution

Finding an appropriate **compromise between local and distributed approaches** forms probably the most important issue. Maintaining local copies of data benefits from better conditions for efficient query evaluation, however, we are not always able or allowed to gather the data under our control.

Scalability

Although some of the existing approaches are able to process large sets of data, experiments performed using various queries and implementations imply that we are still not able to sufficiently flatten the performance of such approaches and the **explosion of the Web of Data size**.

Dynamicity

Data on the Web significantly tend to aging. We especially need not only to handle simple **data modifications**, but also deal with **broken links** and attempt to anticipate or correct them. Unfortunately, the problem is that **index structures are often static** and do not allow any dynamic behaviour.

Quality

The increasing number of globally available data on the Web also causes issues of **data quality, provenance and trust**. Especially in the context of global search engines we need to propose accurate metrics for determining relevance of query results and limiting their number as well.

Model

… what is the purpose and important features of all main notions on which our querying model is built on?

Sources

Sources provide two important functionalities. First, they represent or **contain distributed data** we want to work with. Second, they provide **public interfaces for querying**.

These two components may play very different roles. For example, having an **ordinary flat file** with RDF triples as the simplest form of a source, the interface is only limited to direct access. On the other hand, **SPARQL endpoints** as another example, provide advanced querying support.

Collections

Data triples provided by individual sources are encapsulated into collections. Formally, a **collection is a set of triples**. So, the only problem is to specify which triples of the source belong to a particular collection.

For this purpose we can **select triples explicitly**, or use **descriptions** as a more convenient mechanism allowing us to describe whole groups of triples conforming to required conditions or rules.

Each collection definition may also contain **capabilities** describing query evaluation abilities of a given source and auxiliary **statistics** further helping this query processing.

Collection definitions are maintained by sources themselves and are created by their users. When a source does not follow our model, we can view it as a single collection.

Databases

A **database is a set of collections**, which are generally spread across more distributed and different sources. As a consequence, a database contains all data triples that are associated to all its collections.

The **database definition is created according to our requirements** and is maintained by some source we control.

Storages

Data in sources need not to be stored and maintained directly by these sources. For this purpose we can use existing **native approaches for storing RDF triples**, as well as standard **relational databases** with appropriate wrappers.

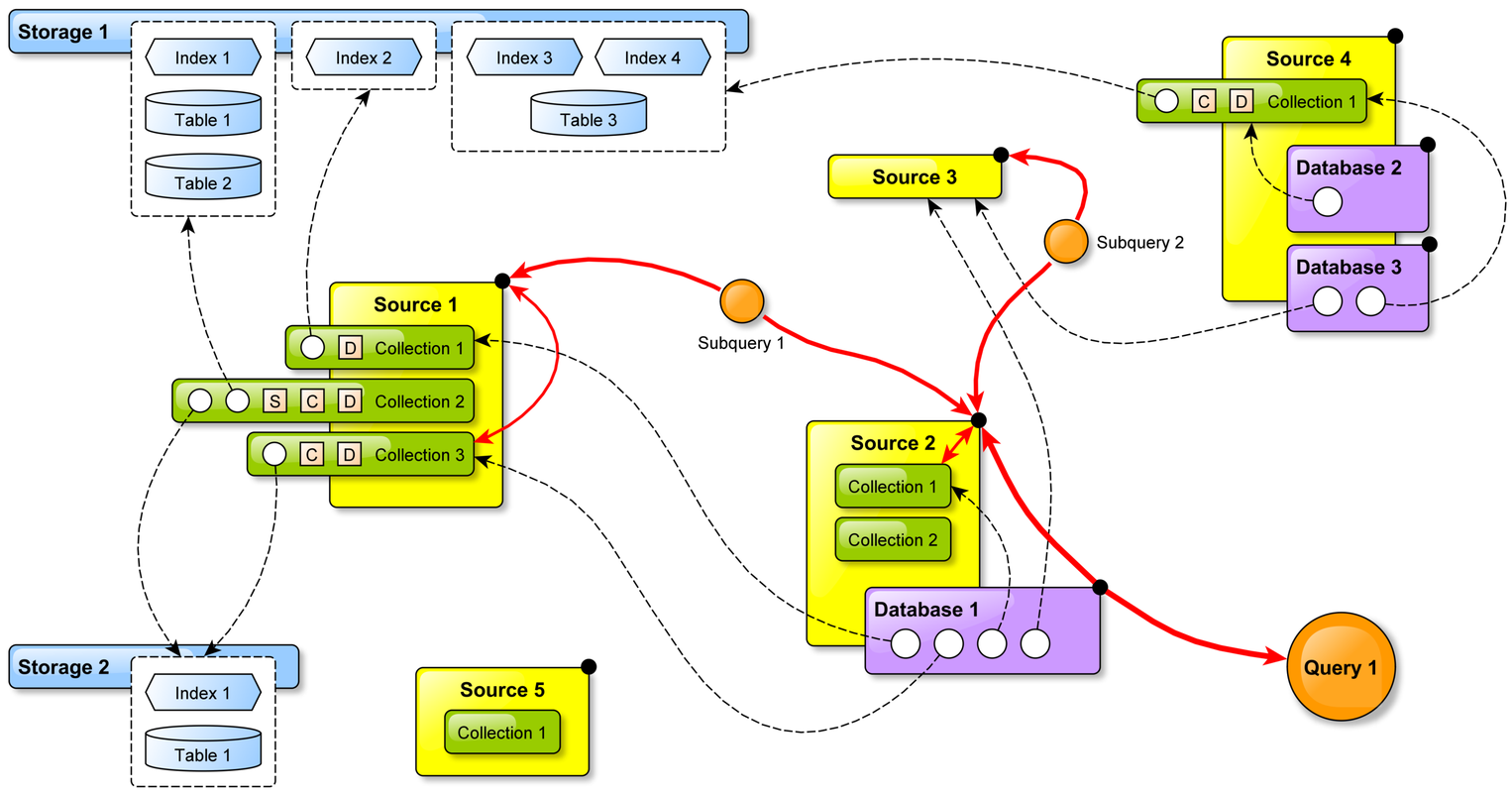
Anyway, each storage may contain data of several collections, even from different sources. Despite **tables** or other structures for triples themselves, they contain **indices** supporting efficient query evaluation.

Queries

Having a **query to be evaluated with respect to a given database**, we access the source containing its definition and use its interface to initiate the processing. With the knowledge of the database definition, this **query is decomposed into subqueries** that are passed to relevant sources and evaluated by them. After this distributed processing is finished, the final results is composed and returned to the user.

Example

… how the query evaluation works on a sample configuration of databases over sources and their collections?



Issues

… which particular issues do we need to figure out?

- **Framework model** — although the main aspects of the system are already sketched in this poster, we still need to deepen the insight into its architecture and design.
- **Data descriptions** — proposal of a model and mechanisms for describing structural and other characteristics of data that are provided in collections by individual sources. Similarly, we need to discuss capabilities and statistics.
- **Data analyses** — the knowledge of characteristics and other features of real-world data may help proposing more efficient data structures and algorithms.
- **Index structures** — auxiliary structures that describe data of distributed databases and enable efficient querying.
- **Query evaluation** — apparently the most complex problem combing all framework design aspects. Query decomposition, distributed evaluation, query plans and different optimisation strategies are only the essentials.

References

… where to find additional information?

- Bizer, C., Heath, T., Berners-Lee, T.: **Linked Data — The Story so far**. In: International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
- Starka, J., Svoboda, M., Mlynkova, I.: **Analyses of RDF Triples in Sample Datasets**. In: Third International Workshop on Consuming Linked Data (COLD@ISWC 2012). Boston, USA. CEUR-WS (2012) [*Accepted for publication*]
- Svoboda, M., Mlynkova, I.: **Efficient Querying of Distributed Linked Data**. In: Proceedings of the 2011 Joint EDBT/ICDT Ph.D. Workshop, pp. 45–50. ACM, New York (2011)
- Svoboda, M., Mlynkova, I.: **Linked Data Indexing Methods: A Survey**. In: On the Move to Meaningful Internet Systems: OTM 2011 Workshops. pp. 474–483. Springer (2011)
- Svoboda, M., Starka, J., Mlynkova, I.: **On Distributed Querying of Linked Data**. Proceedings of the DATESO 2012 Workshop, pp. 143–150. Czech Republic. CEUR-WS (2012)

RWSS 2012, Vienna, Austria, 3. — 8. 9. 2012

This work was supported by the Charles University Grant Agency grant 4105/2011