Querying Large Web Repositories

Matej MARCOŇÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
marconak@gmail.com

Web is one of the largest sources of information in the World. It is necessary to efficiently store and process these data for their further use. Unfortunately, large amount of information on the Internet achieved level, when we are not capable to process this amount of data on a single machine/server. It is necessary to look for other options and approaches how to process large data. One of solutions to the problem is based on parallel data processing on clusters of computers. The most known parallel solution is programming model MapReduce [1]. The main idea of the model is to hide details about parallelism and to allow programmers focus on data processing.

Because the amount of data is increasing, the effective approach for information search is needed. With these thoughts, an idea of integrating some mechanism for recognizing relationship between information on the Web is coming. From this requirement for better machine processing of information, has become trend of use semantics to the Web [3]. Semantics allow us to create webpages or documents, which are more intended for machine processing. This kind of data are often represented as RDF triplets of subjects, predicates and objects (e.g., John, is friend, Mathew) and organized in ontologies. Ontologies and RDF data are standardly queried by SPARQL and its extended forms.

Main objective of our work is to explore the possibility of SPARQL query language and MapReduce techniques for purpose of querying big RDF repositories. It is very important to obtain required information from large RDF repositories as quick as possible. To retrieve the data quickly, it is important to choose a suitable data storage. Therefore we store domain specific data in NoSQL database MongoDB, because data structure aspect is more preferable for our use. Next factor to retrieve information is advanced features of SPARQL. But NoSQL databases do not support querying by SPARQL, so we decided to propose MapReduce algorithm for evaluation of SPARQL and its advanced features.

Most of the existing solutions for co-operation of SPARQL and MapReduce are focused on optimizing graph pattern, but our main goal is an optimal strategy for the implementation of SPARQL's advanced features. In the implementation of advanced

* Supervisor: Karol Rástočný, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 79–80.

features we are going to use different techniques on different levels of programming model MapReduce:

- function *Map*
- function *Finalize*
- combination of above two methods

Implementation of our approach within function *Reduce* is not suitable, due to the functionality of this function is not appropriate for our use. In function *Map*, we can reduce amount of data, that are required for further processing, but unfortunately this phase is restricted only for executing simple operations.

On the other hand, function *Finalize* processes data to final state and in contrast with Map is independent from SPARQL operators. We would like to connect benefits of both methods for better performance.

One of the most used operators in SPARQL is *Filter* and his implementation from the view of its functionality is appropriate for our work [2]. Expressions in operators of Filter can be preprocessed on smaller parts for its application in function Map, due to decreasing amount of data for further processing. In next step in function Finalize is not needed so difficult evaluation of the operators.

We will evaluate our method on the domain specific data in NoSQL database MongoDB. Testing will be conducted on a sample of non-trivial SPARQL queries over this data. Number of this data in MongoDB will be gradually increased for comparing efficiency of our method on different levels of programming model MapReduce.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0233-10.

References

[1] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, (2008), vol. 51, no. 1, pp. 107–113.

[2] Picalausa, F., Vansummeren, S.: What are real SPARQL queries like? In: *Proc. of the International Workshop on Semantic Web Information Management - SWIM '11*. ACM Press, New York (2011). p. 6.

[3] Shadbolt, N., et al.: The Semantic Web Revisited. *IEEE Intelligent Systems*, (2006), vol. 21, no. 3, pp. 96–101.