"Storing Big Data using Natural Language Graph Information Bases"

Krassimir Markov

Abstract:

The term **Big Data** applies to information that can't be processed or analyzed using traditional processes or tools. Three main characteristics define Big Data: ***Volume, Variety, and Velocity*** [Zikopoulos et al, 2012].

These characteristics cause corresponded **problems of storing Big Data**. Big Data created the need for a new class of capabilities to augment the way things are done today to provide better line of site and controls over our existing knowledge domains and the ability to act on them.

Popular approach for representing Big Data is **Resource Definition Framework (RDF)** as well as several equivalent approaches, such as XML and OWL. Let remember, RDF is a graph based data format which is schema-less, thus unstructured, and self-describing, meaning that graph labels within the graph describe the data itself. The prevalence of RDF data is due to variety of underlying graph based models, i.e. almost any type of data can be expressed in this format including relational and XML data [Faye et al, 2012].

The state of the art with respect to existing storage and retrieval technologies for RDF graphs is given in [Hertel et al, 2009]. Different repositories are imaginable, e.g. main memory, files or databases. RDF schemas and instances can be efficiently accessed and manipulated in main memory. For persistent storage the data can be serialized to files, but, for large amounts of data, the use of a database management system is more reasonable. Examining currently existing RDF stores we found that they are using relational and object-relational database management systems. Storing RDF data in a relational database requires an appropriate table design.

Graph database models took off in the eighties and early nineties alongside object oriented models. Their influence gradually died out with the emergence of other database models, in particular geographical, spatial, semi-structured, and XML. Recently, the need to manage information with graph-like nature has reestablished the relevance of this area [Angles & Gutierrez, 2008].

The graph oriented approach for storing ontologies became one of the preferred. Some of the world's leading companies and products which support extra-large ontology bases are presented on page of W3C [LTS, 2012]. It should be noted, there exists a gradual transition from relational to non-relational models for organizing ontological data. Perhaps the most telling example is the system AllegroGraph® 4.9 [AlegroGraph, 2012] of the FRANZ Inc.

In the Big Data community, the **"MapReduce Paradigm"** has been seen as one of the key enabling approaches for meeting the continuously increasing demands on computing resources imposed by massive data sets. MapReduce is a highly scalable programming paradigm capable of processing massive volumes of data by means of parallel execution on a large number of commodity computing nodes. It was recently popularized by Google [Dean & Ghemawat, 2008], but today the MapReduce paradigm has been implemented in many open source projects, the most prominent being the

Apache Hadoop [Hadoop, 2014]. The popularity of MapReduce can be accredited to its high scalability, fault-tolerance, simplicity and independence from the programming language or the data storage system.

At the same time, **MapReduce faces a number of obstacles** when dealing with Big Data including the lack of a high-level language such as SQL, challenges in implementing iterative algorithms, support for iterative ad-hoc data exploration, and stream processing [Grolinger et al, 2014].

A possible solution may be the **Collect/Report Paradigm** and Natural Language Addressing approach (NL-addressing). It is suitable for storing Big Data in large information bases located on different storage systems – from personal computers up to cloud servers [Markov et al, 2015].

NL-addressing consists in assuming the internal computer codes of letters as co-ordinates in multi-dimensional information space. Different words and phrases have different lengths and require using of addressing with variable length of the co-ordinate arrays, i.e. to have variable dimensions in one and the same time. Such addressing we call "**multidimensional**".

Our starting point of realization of our approach is the **Multi-Domain Information Model (MDIM)** [Markov, 2004] and corresponded **Multi-Domain Access Method (MDAM)** [Markov, 1984], which we upgraded to NL-addressing approach to apply for storing graphs. The possibility to use coordinates is good for graph models where it is possible to replace search with addressing. Hence, the advantages of the numbered information spaces are:

- The possibility to build growing space hierarchies of information elements;
- The great power for building interconnections between information elements stored in the information base;
- The practically unlimited number of dimensions (this is the main advantage of the numbered information spaces for graphs where it is possible "*to address, not to search*");

The NL-addressing and multi-layer organization of the information, together with the model of representing the characteristics, are good basis for implementing this approach for real solutions [Markov et al, 2015].

Bibliography

[AlegroGraph, 2012] AllegroGraph® 4.8, http://www.franz.com/agraph/allegrograph/ (accessed: 25.08.2012).

[Angles & Gutierrez, 2008] Angles R., C. Gutierrez, "Survey of Graph Database Models", ACM Computing Surveys, Vol. 40, No. 1, Article 1, Publication date: February 2008, DOI 10.1145/1322432.1322433, http://doi.acm.org/10.1145/1322432.1322433, pp. 1-39

[Dean & Ghemawat, 2008] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, 51(1), 2008, pp. 107-113.

[Faye et al, 2012] David C. Faye, Olivier Cure, Guillaume Blin, "A survey of RDF storage approaches", Received, December 12, 2011, Accepted, February 7, 2012, ARIMA Journal, vol. 15, 2012, pp. 11-35.

[Grolinger et al, 2014] K. Grolinger, M. Hayes, W. Higashino, A. L'Heureux, D. S. Allison, M. A. M. Capretz, “Challenges for MapReduce in Big Data”, Proc. of the IEEE 10th 2014 World Congress on Services (SERVICES 2014), Alaska, USA, June 27-July 2, 2014

[Hadoop, 2014] Apache Hadoop, http://hadoop.apache.org . (accessed: 22.12.14)

[Hertel et al, 2009] Hertel A., J. Broekstra, and H. Stuckenschmidt, “RDF Storage and Retrieval Systems”, In: S. Staab and R. Studer (eds.), Handbook on Ontologies, International Handbooks on Information Systems, DOI 10.1007/978-3-540-92673-3, Springer-Verlag Berlin Heidelberg 2009. pp 489-508.

[LTS, 2012] LargeTripleStores http://www.w3.org/wiki/LargeTripleStores (accessed: 29.08.2012)

[Markov et al, 2015] Krassimir Markov, Krassimira Ivanova, Koen Vanhoof, Vitalii Velychko, Juan Castellanos, „Natural Language Addressing”, ITHEA® Hasselt, Kyiv, Madrid, Sofia, IBS ISC No.: 33, 2015, ISBN: 978-954-16-0070-2 (printed), ISBN: 978-954-16-0071-9 (online), 315 p.

[Markov, 1984] Markov Kr. A Multi-domain Access Method.//Proceedings of the International Conference on Computer Based Scientific Research, Plovdiv, 1984. pp. 558-563.

[Markov, 2004] Markov, K. Multi-domain information model, Int. J. Information Theories and Applications, 11/4, 2004, pp. 303-308.

[Zikopoulos et al, 2012] Paul C. Zikopoulos, Chris Eaton, Dirk de Roos, Thomas Deutsch, George Lapis, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, Copyright© 2012 by The McGraw-Hill Companies, ISBN 978-0-07-179053-6, MHID 0-07-179053-5, 2012, 166 p.