International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Survey Paper on Technologies to Link and Model Multidimensional Data on the Semantic Web for Business Intelligence

Karan Gupta[1], Prof. Poonam D. Lambhate[2]

[1]M.E (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India -411007

[2]Professor, Department of IT and Computer Engineering, Jayawantrao Sawant College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India -411007

Abstract: ***The vast amount of data on the World Wide Web has resulted in convergence powerful analytical technologies, namely OLAP and the Semantic Web. The Semantic Web has allowed for evolution of linked multidimensional data that allows for querying information on the web using their semantics or meaning and not just a list of key words. Business Intelligence platforms provide great support in analyzing large volumes of data. This has been further extended to map and model multidimensional data on the Semantic Web in the form of OLAP cubes which are expressed in RDF vocabularies. This paper surveys the current technologies that provide a platform for building and querying data warehouses on the semantic web using multidimensional OLAP data cubes built using the RDF data cube vocabulary. The aim of this paper is to present a roadmap for new developments and research in the field of Semantic Web.***

Keywords: Linked Multidimensional Data, OLAP, RDF, Semantic Web

1. Introduction

The Semantic Web [1] has grown to be a great platform to share and build information on the World Wide Web. The current standards for HTML and CSS only allow for displaying and positioning of data in an aesthetic manner, but there is no inherent way of also preserving the meaning of the data. The semantic web as the name suggests allows for processing web data using the semantics or meaning of the data. It does this by using a collection of technologies [2] like Extensible Markup Language (XML), Resource Description Framework (RDF), RDF Schema and Web Ontology Language (OWL). These collectively provide a platform for applications to query the data on the web and draw inferences using RDF vocabularies. Data on the semantic web can be inherently linked based on their common origins. However, this can only be possible if all the data on the web is available in a format that is standardized for use by the various technologies. The semantic web aims at creating data that can be linked using relationships just like a relations database. Once data can be linked it can provide for true conversation between the vast amount of heterogeneous and unstructured data on the web.

2. Semantic Web Overview

The semantic web achieves linking of the data using Vocabularies, also referred to as Ontologies. The primary role of Ontologies is to help with integration of data, remove ambiguities and to organize knowledge. “An ontology is defined as a formal, explicit specification of a shared conceptualization”. The architecture of the Semantic Web [1] is illustrated in Figure 1.

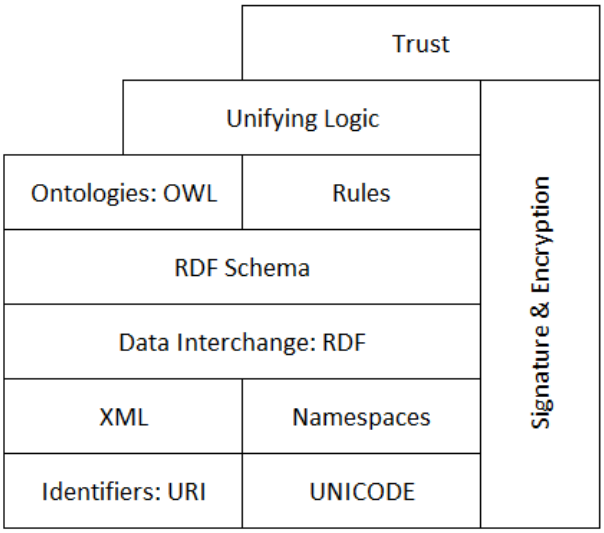

Figure 1: Semantic Web Architecture

The extensible markup language allows for proper syntax to model document contents. RDF [3] gives a graph based model to describe objects and their relationships. RDF Schema [4] provides support for a vocabulary and axioms for describing properties and classes of the RDF-based resources. OWL [5] adds additional vocabulary support for describing properties and classes in order to construct ontology.

Compared to traditional web technologies which focus mainly on representing data, the Semantic Web provides for a more machine-readable platform that allows for the extraction of information about web resources and relationships between various heterogeneous resources.

The primary role of the Semantic Web technologies is to define a common vocabulary of standard and constraints (inferring rules) in order to create a semantic metadata. The semantic web data should follow four principles [6]:

- Use of Uniform Resource Identifiers (URIs) to identify object.
- Use of Hypertext Transfer Protocol (HTTP) to facilitate searching for objects.
- Us of se the Resource Description Framework (RDF) [3] format as a standard to provide descriptive information about an object

3. Literature Survey and Related Work

In spite of its many advantages, representing multidimensional data on the Semantic Web is still in its early stages. OLAP was conceived primarily to provide analysis over homogenous sets of data. A lot of research has been done to provide OLAP functionality on heterogeneous and unstructured data and to link this to the Semantic Web.

In 2007, Romero et Abelló [7] proposed an approach to define an OLAP data warehouse from a single domain ontology. It was semi-automatic and the OLAP warehouse had the potential to be integrated with unstructured web sources. This approach paved the way for OLAP analysis to be carried out on Semantic Web data. However, there were many limitation s of this approach as it is limited to single domain ontology. Information on the web spans across several domain ontologies, it cannot cater to overlapping entities from multiple domain ontologies.

In 2009, Nebot et al. [8] propose a new framework to define semi-structured data warehouse from multiple domain ontologies. This data warehouse was called Semantic Data Warehouse (SDW) and used ontology mappings in order to manage domain overlaps.

There was also significant research undertaken to combine advantages of Linked Data Platform (LDP) Vocabulary [18] and RESTful data approach.

Alarcon et al. [19], in 2010, presented a framework on building RESTful SPARQL mappings. They propose a redesign of SPARQL to allow for the evolution of the Semantic Web in a decentralized manner. They also explored limitations of the mapping SPARQL to the corresponding HTTP methods.

Harth et al. [20], in 2011, proposed Linked Data Services (LIDS), an approach to integrate data services with Linked Data. Their algorithms allowed for the automatic creation of links between data sets and services

These and many other research works still had the limitations of storing the extracted Semantic Web data into a local OLAP data warehouse cube. As these approaches were semi-automatic, it was infeasible to cater to the dynamically changing nature of the web and extract data in real-time.

Any OLAP analysis would require expensive and complex ETL processes to extract and load the Semantic Web data. Research then shifted to carry out OLAP analysis directly on the Semantic Web without any ETL processes. This led to evolution of frameworks based on the RDF Data Cube Vocabulary (QB) [9]. Most of the modern day frameworks are variants of QB. (Fig 2) [9].

QB, however did not allow representation of hierarchal dimensional data across multiple levels, was still not capable of providing true querying capabilities on Semantic OLAP cubes.

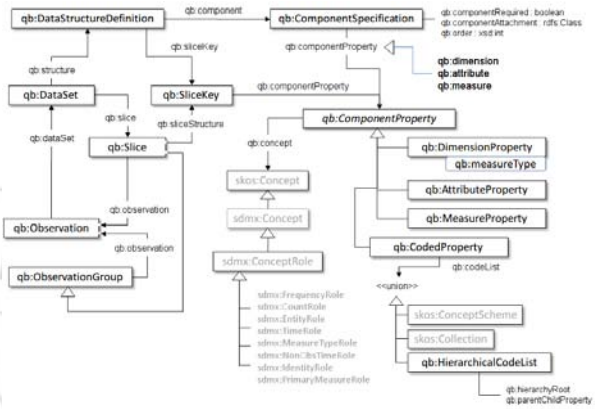

Figure 2: A pictorial representation of the QB Vocabulary

In 2011, Kämpgen et al [10] extended the QB model to represent statistical data in a multidimensional format. They presented a model to map statistical linked data that conformed to the QB vocabulary. In 2012, they demonstrated through the use of SPARQL [11] on how to generate facts from the cube through nested OLAP operations.

Harth et al. [21], in 2012, extended Linked Data with RESTful services. They proposed a system that enables data-driven application development built using the RESTful use of Linked Data.

Researchers again concentrated on addressing the limitations of the RDF Data Cube Vocabulary.

In 2012, Etcheverry et al. [12] introduced the Open Cubes Vocabulary (OC), new multidimensional language that supported multiple dimension hierarchies. However, the major limitation was the OC was not compatible with previous QB applications. This was because OC was a specific modeling language and did not allow for reusing data already published using QB.

Later in 2012, Etcheverry et al. [13] proposed a new vocabulary that extended QB to fully support OLAP models and operators. This new vocabulary was called QB4OLAP. They provided algorithms to transform cubes based on QB into an equivalent QB4OLAP cube, thereby overcoming the limitations of the Open Cube Vocabulary approach. QB4OLAP allowed for multidimensional modelling of Semantic Web data.

In the same year, Salas et al. [14], presented a plugin called OLAP2DataCube for Ontowiki that transformed a relational star schema into an OLAP cube.

In 2013, Saad et al. [15] provided a framework modelled using QB that supported multi-dimensional and fact analysis with multiple hierarchies. They also made extensive use of SPARQL to implement the various OLAP operations.

There was also related research work in the triplication of multidimensional data.

In 2013, Ruback et al. [16] presented mediation architecture to help describe and consume statistical data, exposed as RDF triples, but stored in relational databases. They propose that a catalogue is just the description of data cube and excludes facts and is hence, not a complete materialization of the relational database.

In 2014, Etcheverry et al. [17] presented an extension to the QB4OLAP vocabulary to support aggregate table and a framework to translate an existing relational warehouse into its equivalent schema QB4OLAP. This also allowed for possibilities of using SPARQL to query the QB4OLAP schema. Their algorithms allowed to translate OLAP operations such as Roll-Up and Slice to SPARQL queries. (Figure 3)

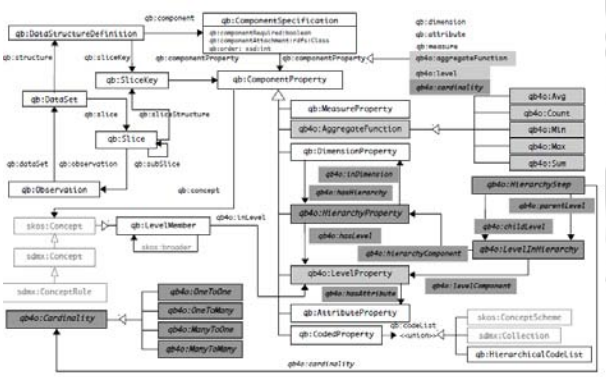

Figure 3: A pictorial representation of the QB4OLAP Vocabulary

4. Proposed Work

In this survey we looked at various technologies that allow to link open multidimensional data which can be later used for querying. SPARQL [22] a recursive acronym for SPARQL Protocol and RDF Query Language) is an RDF query language, that is, a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. This can be used to query multiple linked data cubes over the Semantic Web. A conceptual framework is proposed that can allow for publishing semantically linked data using RDF and then combine all such web cubes using commonalities between them. A RDF browser can be built that uses SPARQL and query the underlying cube built using QB4OLAP. There are several limitations in the existing data cube vocabularies like performance and querying over real-time data. Introducing indexes in the data cubes can help overcome these issues. In order to achieve this, the existing RDF Data Cube Vocabulary will need to be extended further

5. Conclusion

There are still various challenges that need to be overcome to effectively combine OLAP and the Semantic Web. The most common way to map unstructured (or semi structured) data in OLAP data cube is to create mappings through ontology. However, building the ontology manually is an extremely complex task. Many ongoing researches are now looking at building the ontology automatically and create mapping between heterogeneous data sources This paper provides an up-to-date overview of researches that aim to enhance OLAP analysis in the BI field with SW technologies. It discussed how Semantic Web technologies evolved from the traditional ETL based to the more automatic mapping of multidimensional data for OLAP analysis using the RDF Data Cube Vocabulary.

References

[1] T. Berners-Lee, J, Hendler, O. Lassila, “The Semantic Web. Scientific American” 284(5), pp. 35–43, 2001.
[2] http://www.w3.org/standards/semanticweb/data
[3] K. Graham, C. Jeremy, “Resource Description Framework (RDF): Concepts and Abstract Syntax”, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, 2004.
[4] Brickley Dan, Guha RV. RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/; 2004
[5] D. Mike, G. Schreiber, “OWLWeb Ontology Language Reference”, http://www.w3.org/TR/2004/REC-owl-ref-20040210/, 2004.
[6] http://www.w3.org/DesignIssues/LinkedData.html.
[7] Romero, Oscar, et Alberto Abelló, “Automating multidimensional design from ontologies. ACM Press, p. 1‐ 8, 2007.
[8] Nebot, Victoria, Rafael Berlanga, Juan Manuel Pérez, María José Aramburu, et Torben Bach Pedersen. “Multidimensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses”. Springer Berlin Heidelberg, p. 1‐ 36, 2009.
[9] http://www.w3.org/TR/vocab-data-cube.
[10] Kämpgen, Benedikt, Sean O’Riain, et Andreas Harth. 2012. “Interacting with statistical linked data via olap operations”, 2012.
[11] http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.
[12] Etcheverry, Lorena, et Alejandro A. Vaisman “Enhancing OLAP Analysis with Web Cubes” Springer, 2012.
[13] Etcheverry, Lorena, et Alejandro Vaisman. QB4OLAP: A New Vocabulary for OLAP Cubes on the Semantic Web. Springer, 2012.
[14] Percy E. Rivera Salas, Michael Martin, Fernando Maia Da Mota, Sören Auer, Karin K. Breitman, and Marco A. Casanova. 2012. OLAP2DataCube: an ontowiki plug-in for statistical data publishing.

[15] Saad, Rafik, Olivier Teste, et Cássia Trojahn. "OLAP Manipulations on RDF Data following a Constellation Model", 2013.
[16] Lívia Ruback, Marcia Pesce, Sofia Manso, Sérgio Ortiga, Percy E. Rivera Salas, and Marco A. Casanova. A mediator for statistical linked data. ACM, 2013
[17] Etcheverry, Lorena, Alejandro Vaisman, et Esteban Zimányi. "Modeling and Querying Data Warehouses on the Semantic Web Using QB4OLAP". Springer, 2014.
[18] http://www.w3.org/ns/ldp
[19] Alarcón, R., Wilde, E.: From restful services to rdf: Connecting the web and the Semantic Web. California, 2010.
[20] Sebastian Speiser, Andreas Harth, "Integrating Linked Data and Services with Linked Data Services". Proceedings of the 8th Extended Semantic Web Conference, Springer.
[21] Steffen Stadtmüller, Andreas Harth. "Towards Data-driven Programming for RESTful Linked Data". Proceedings of the ISWC 2012 workshop on Programming the Semantic Web.
[22] Etcheverry, Lorena, Alejandro Vaisman, et Esteban Zimányi. "Modeling and Querying Data Warehouses on the Semantic Web Using QB4OLAP". Springer, 2014.

Author Profile

Mr. Karan Gupta is currently pursuing M.E (Computer) from Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, Savitribai Phule Pune University, Pune, Maharashtra, India-411007. He received his B.E (Computer) Degree from JSPM's, Jayawantrao Sawant College of Engineering, Pune, Savitribai Phule Pune University, Pune, Maharashtra, India -411007. His area of interest includes Business Intelligence and analytics over the Semantic Web.

Prof. Poonam Lambhate received the B.E. and M.E. degrees in Computer Engineering from Solapur University and Shivaji University, India. She is currently associated with the department of Information Technology Department at JSPM's Jayawantrao Sawant College of Engineering, Savitribai Phule Pune University, India. Her research interests include Image processing, Information retrieval, swarm intelligence; data mining. She is also a member of ISTE.