# Scalable methods to analyze Semantic Web data

Victoria Nebot Romero
*Lenguajes y Sistemas Informáticos, Universitat Jaume I, 12071, Castellón, Spain, E-mail: romerom@uji.es*

## 1. Introduction

Semantic Web data [2] is currently being heavily used as a data representation format in scientific communities, social networks, business, news portals and other domains. The irruption and availability of Semantic Web data is demanding new methods and tools to efficiently analyze such data and take advantage of the underlying semantics. Although there exist some applications that make use of Semantic Web data, advanced analytical tools are still lacking, preventing the user from exploiting the attached semantics.

The main objective of this dissertation is to provide a formal framework that enables the multidimensional analysis of Semantic Web data in an scalable and efficient manner [5]. The success of multidimensional analysis techniques applied to large volumes of structured data in the context of business intelligence, especially for data warehousing and OLAP applications [3,4], has prompted us to investigate the application of such techniques to Semantic Web data, whose nature is semi-structured and contain implicit knowledge.

## 2. Motivating scenario

In a few words, the Semantic Web is an extension of the current web in which information is given a well-defined meaning. This is achieved by annotating data with structured and machine-processable metadata. Resources are assigned a well-defined meaning and interpreted by means of ontologies, which encode knowledge about a particular domain.

Imagine the healthcare scenario, where patient's data across different hospitals are semantically annotated, integrated and linked to biomedical data using a semantic framework. That is, patients' diseases, symptoms, treatments, etc. are correctly identified and linked to concepts in ontologies, where their meaning and taxonomical relations are specified. Also, patient's measurements such as blood pressure or heart rate are not mere numbers but have a well-defined meaning. Such a semantic framework where data are leveraged to a conceptual level would enable more in deep and thorough analysis of patients' data as well as meaningful "dashboards" for well-informed decision-making.

Fortunately, the previous scenario is a reality. The amount of Semantic Web data is growing exponentially, especially in complex scenarios such as healthcare or Life Sciences, where semantics alleviate integration issues. However, traditional analytical tools for decision making based on multidimensional modeling are not appropriate for data that is complex, semi-structured and dynamic.

The multidimensinal model has been successfully applied to traditional decision support because of its simplicity. It is based on the fact/dimension dichotomy. Data are modeled in terms of facts, or observations, which contain analytical measures, and dimensions, which are the different analysis perspectives. Dimensions are usually hierarchically organized so that facts can be summarized at different granularity levels (dimension levels) by applying a summarization function to the measures.

Even though some attempts have been made to enable multidimensional analysis beyond relational data to semi-structured data [1], the truth is that none of the approaches have addressed the problem with Semantic Web data in all its complexity.

This thesis proposes a multidimensional analytical framework suitable for Semantic Web data that takes advantage of all the implicit knowledge to make more meaningful and sophisticated analysis.

This way, doctors will be able to analyze patients by selecting research variables (or dimensions) from semantically enriched data, such as the disease type, the patient's gender, the type of administered drug, etc. in order to analyze the impact on several indicators (or measures), such as clinical tests results or disease recovery levels. By enabling all the capabilities of multidimensional analysis over Semantic Web data, doctors

XX
ISSN XX, IOS Press. 

will be able to create summaries at different granularity levels. For example, they will be able to display average cholesterol levels of patients with cardiovascular diseases, and right on, zoom in to display only the measurements of patients having an angina pectoris. This analysis will be possible thanks to the semantic relations between disease types encoded in the ontologies, which indicate that an angina pectoris is a type of cardiovascular disease. This knowledge is usually implicit in domain ontologies and we will exploit it to perform multidimensional analysis.

## 3. Methodology

The proposed analytical framework addresses all the challenges related to the manipulation, processing and analysis of Semantic Web data to enable efficient, scalable and full-fledged multidimensional analysis.

Scalability is achieved by two means. On one hand, we provide an ontology indexing model over ontologies that allows to manage implicit knowledge in a compact format. Therefore, operations that require reasoning can be efficiently solved using the indexes. On the other hand, we have developed several scalable modularization techniques that build upon the previous indexes and allow to extract and work only with the ontological subsets of interest.

These methods are used to make the extraction of facts and dimensions from Semantic Web data efficient and scalable. However, identifying facts, dimensions, measures and well-shaped dimension hierarchies from the graph structure that underlies SW data is a big challenge due to the mismatch between the graph model that underlies Semantic Web data and the multidimensional model. Therefore, the notions of fact and dimension are revisited in the Semantic Web context and both facts and dimensions are defined from a logical viewpoint.

Facts are formally defined as multidimensional points quantified by measure values, all of which are logically reachable from the subject of analysis defined by the user. To this end, we use the notion of aggregation path, which is less restrictive than the traditional multidimensional constraints usually imposed between facts and dimensions, and define different interesting subgroups for analysis. We also detect the summarizability of the extracted facts and produce correctly aggregated results.

Dimensions are defined as direct acyclic graphs composed by nodes that are semantically related and adhere to the conceptual specification of the user. That is, nodes are sub-concepts of the dimension type and the edges are subsumption relations.

The flexibility introduced in the discovery of facts and dimensions allows us to analyze data that have complex relations, which escape to the traditional multidimensional model and cannot be otherwise analyzed. Moreover, instead of building a huge, one time data warehouse, our method for fact and dimension extraction is based on several indexes and precomputed data, which allows us to efficiently materialize only the facts and dimensions required by an analytical query. This provides up-to-date, dynamic and customized results to the user.

## 4. Conclusion and future work

The analysis framework developed in the dissertation constitutes a powerful tool that can be used in several domains (healthcare, business, etc.) to analyze semantic data in sophisticated ways until now unseen. The developed use case shows the analytical potential of the framework and the experiments validate the viability of the proposal and demonstrate the efficiency of the developed methods.

There are several lines of future work that involve the enhancement of the indexes, the inclusion of properties in the modularization approaches, the definition of new types of aggregation paths, among others. In a broader sense, the success of the developed methods encourages us to widen the analysis perspectives on Semantic Web data to other kinds of analysis that aid decision making, specially data mining techniques.

Advisor:
Prof. Dr. Rafael Berlanga

## References

[1] R. Berlanga, O. Romero, A. Simitsis, V. Nebot, T. B. Pedersen, A. Abelló, and M. J. Aramburu. Semantic Web Technologies for Business Intelligence. pages 310–339, 2012.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[3] E. F. Codd, S. B. Codd, and C. T. Salley. Providing OLAP (On-Line Analytical Processing) to User Analysts: An IT Mandate. E. F. Codd and Ass., 1993.

[4] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley, 2nd edition, April 2002.

[5] V. Nebot. *Scalable methods to analyze Semantic Web data*. PhD thesis, Universitat Jaume I, Castellón, Spain, 2013. http://maat.dlsi.uji.es:8000/tkbgtest/PhDs/PhDVic.pdf/view.