Graduate Research in Engineering and Technology (GRET)

Volume 1 | Issue 1 Article 8

July 2013

WEB SEMANTICS DATA

M. MARKANDEYULU
Department of CSE, Vignan's Lara Institute of Technology & Science, Guntur,India,
Markyellow9@gmail.com

CH. JAYARAJU
Department of CSE, Vignan's Lara Institute of Technology & Science, Guntur,India, Glorious.ch@gmail.com

WEB SEMANTICS DATA

M.MARKANDEYULU[1] & CH.JAYARAJU[2]

[1,2]Department of CSE, Vignan's Lara Institute of Technology & Science, Guntur,India
E-mail: Markyellow9@gmail.com, Glorious.ch@@gmail.com

Abstract- Data Semantics is a wide area that continuously faces new challenges arising from the invention of new information formats and novel applications. An area that is partic- ularly challenging with respect to identifying, representing and using data semantics is the Web. This paper attempts to characterize the nature and challenges of Data Semantics on the Web as an interesting research area to be covered by the Journal on Data Semantics.

I. INTRODUCTION

Data Semantics is a topic that has been investigated in computer science for more than 30 years. It is typically associated with a formal definition of the intended interpre- tation of the data often in terms of logic or algebraic formal- isms .Over the time, the goals of defining data semantics as well as the ideal of having a clear formal representation of semantics has not change, what has changed and is con- stantly changing, however, are ways of capturing and using the semantic of data as well as the formalisms used to rep- resent it. These changes are triggered by new kinds of appli- cations that require new types of data (e.g. geoinformation or social tagging) and with advances in the start of the art of data management (e.g. distribution and parallel process- ing) that has come with new problem with respect to data semantics.

One invention that brought significant changes to the field of data semantics is the Web. The Web fundamen- tally changed the way data is managed if compared to tradi- tional systems. In traditional systems, the basic idea is to keep the system in a consistent state or to move it from one consistent state into another. On the web, many traditional assumptions of data semantics are not valid any more. This makes it hard to even define the notion of a consistent state. As a consequence, data semantics on the web require new methods and principles to be developed. As an answer to this need, a rather active research filed has evolved on such principles and method within the broader area of semantic web technologies. The goal of this research is roughly to transfer traditional ideas of data semantics to the area of web data taking into account the specific challenges and needs of a data infrastructure like the web.In this paper, I try to define the research area of data semantics on the web by giving an overview over the challenges and ideas in that part of semantic web research concerned with data semantics in a wider sense. We start by discussing the nature of data on the web focussing on the kind of data typically found on the web and specific chal- lenges we have to face in the area of Data Semantics. I argue that the web does not come with challenges that are unique, but that we are facing a unique combination of challenges that establishes Data Semantics on the Web as a research area in its own rights. After this general discus- sion of the challenges I discuss three basic aspects of Data Semantics on the Web, namely the extraction of semantics from Web data, the representation of the semantic Informa- tion on the Web and the Use of such Semantic represen- tations for processing data on the Web. I will try to stay away from specific solutions and applications as much as possible and rather focus on general principles and lines of work. I will conclude with a personal view on impor- tant research directions in the field of Data Semantics on the Web that need further attention in future research in this area.

II. THE NATURE OF WEB DATA

Web Data comes in a variety of forms that have emerged along with the development of the web over the past 20 years. While each of these different types of data have their own characteristics and requires different kinds of processing methods and infrastructures, they also have a lot of com- monalities in terms of challenges for data processing.

2.1 Types of Data on the Web

During te past 20 years, the web has evolved from a doc- ument management system used internally at CERN into a global information medium that becomes more important for all parts of the society including science, business, politics and social relations. This development has taken place in a number of phases that can roughly be associated with differ- ent kinds of data shared over the Web. We briefly recall these developments and the different kinds of Web data. *Documents and Web Pages* Initially, the Web was created as a hypertext system for sharing research results in terms of manually created HTML pages and documents that are linked to these pages .This phase is sometimes referred to a Web 1.0. While today, web pages can contain many

dif- ferent kinds of multimedia information, the dominant kind of information on web pages is still in the form of natural language texts. Thus, the semantics of web data is to some extend always connected to natural language semantics and being able to process natural language resources is a basic requirement for semantic processing of web data.

Documents and Web Pages Initially, the Web was created as a hypertext system for sharing research results in terms of manually created HTML pages and documents that are linked to these pages .This phase is sometimes referred to a Web 1.0. While today, web pages can contain many dif- ferent kinds of multimedia information, the dominant kind of information on web pages is still in the form of natural language texts. Thus, the semantics of web data is to some extend always connected to natural language semantics and being able to process natural language resources is a basic requirement for semantic processing of web data.

Databases and the Deep Web Although text is still the most visible form of data on the Web, a significant amount of data on the web today is structured data from databases that have been linked to web pages. The resulting information space available through the web is often referred to as the
'Deep Web' or the 'Hidden Web' as the data is typically not explicitly represented on web pages and is thus not indexed by conventional search engines, but has to be surfaced by posting queries to database interfaces on web pages. While the semantics of relational data is well understood, the problem of data on the deep web is the need to describe the data indirectly via the available interface.

Social Media A rather recent development is the spread of so-called social media. While all other kinds of data dis- cussed so far are created by data providers, social media provide data consumers with the possibility to add contents to web pages. This User-created contents has a number of specifics that need to be taken into account. In particular, social contents is characterized by a very low level of regu- larity that is even below the regularity of natural language. Further, user-created content is highly diverse and subjective asking for methods able to control these aspects. In contrast to other forms of data, user created contents is often very timely and therefore provides an important indicator for trends and buzzes.

2.2 Challenges in Web Data Management

While the nature of the different types of data differs, certain challenges for managing data (semantics) on the Web come from the Web itself and are therefore similar for the different data types. These challenges that are described in the follow- ing also define the research area of web data management by setting it aside from classical data management. There is no doubt that similar challenges occur in different areas of data management, the combination of all of the following challenges, however, is rather unique.

Heterogeneity Heterogeneity of data is a fundamental prob- lem on the web. As we have seen above, heterogeneity starts with the problem of having many different kinds of data rep- resenting using a variety of representations including free text, XML languages and relational data. However, hetero- geneity of data on the Web is not limited to data formats and representations, but also occurs at the level of conceptual models and terminology used to describe data items, often referred to as semantic heterogeneity. While semantic heterogeneity is already a major problem when staying inside the relational data model, it becomes a real challenge on the web, where semantic heterogeneity has to be addressed across different types of data.

Change While managing change is a problem in data man- agement in general, it is especially difficult on the web because, as mentioned above, there is no central point of con- trol. Different web sites changes completely independent of each other and there are no mechanisms for propagating or even announcing changes to other sources referring to the changes data. The original approach of coping with change on the web is not to care about it and just ensure that the sys- tem remains stable. While this approach has proven to be very effective, it is problematic from the point of data semantics, especially when the meaning of data in one source depends on the interpretation of another.

Scale Probably, the most significant challenge of seman- tic web data management is scale. The web is the largest freely available information resource that ever existed and it is constantly growing. The following statistics underline this aspect.

III. CREATING DATA SEMANTICS

In closed systems, the intended meaning of data is defined by its intended use that is determined by the systems' develop ers and users and often reflected in the specific schema or the data structures used for representing it. On the web, this is only partially the case. While the intended meaning of data is of course also determined by the intended use, the universal availability of the data via the Web infrastructure encourages the use of data for applications different from the one it was originally intended for. In order to do this in a meaningful way, the intended meaning of the data has to be understood to correctly relate it to the new application. Thus getting hold of the intended meaning, the Semantics of Data on the Web is essential. We can distinguish different general approaches to the problem of understanding the intended meaning of Data on the web.

3.1 Semantics from Models

A viable way of dealing with the problem arising from the attempt to use data for a different purpose than originally intended is to make the intended meaning of the data explicit by publishing the conceptual model in terms of an ontology and linking the data to it using metadata .Meanwhile, this approach is well supported by language standards such as RDF and OWL and can be seen as a cornerstone of data semantics on the web .A closer look at this idea reveals, that publishing the ontology along with a data set does not really solve the problem as long as every data set comes with its own ontology. In this case the problem of possible mis-interpretations is just lifted from the data to the conceptual level. In order to be able to really interpret the data, what is needed is either a jointly used ontology that is shared between the data source and the potential user of the data. In various domains, such ontologies have been developed that can be used to assign an agreed interpretation to data in that domain. Further, some so-called top-level ontologies have been developed that provide a common interpretation at a very high level and can be used to harmonize domain ontol- ogies. The other possible solution is the use of semantic mappings either between the ontologies of data provider and consumer or directly between different data- sets. Meanwhile, it is commonly accepted that semantic mappings between models or data are an important mecha- nism in the description of data semantics, especially on the Web. In particular, the use of same-as links to indicate differ-ent representations of the same real-world object has recently gained a lot of attention in the context of linked data.

IV. DATA SEMANTICS REPRESENTATION

As we have seen, the intended interpretation of Data on the web can be explored in different ways, but it is widely agreed that models of this intended meaning play a central role in intelligent information processing. With the introduction of OWL as a W3C standard for representing ontologies, the representation of data semantics is often associated with it. The spectrum of models for representing Data semantics, however, is actually broader than that. Instead of discussing specific languages that have been proposed for this purpose, we rather discuss different principles underlying these lan-guages that provide different ways of approaching the prob- lem of representing semantics.

This of course includes logic as a classical way of representing meaning, but the charac- teristics of Web Data discussed in Sect. ask for more than a purely logical treatment of data semantics. In particular, successful approaches have to deal with uncertainty and lin- guistic/pragmatic semantics of data on the Web.

4.1 logics

The traditional way of defining Data Semantics in terms of algebraic structures and formal logic using the model theoretic semantics of the respective formalism as a mathematical framework for defining and analyzing the semantics. In databases, Datalog has become the most important model for talking about the semantics of relational data ,in the field of knowledge representation, description logics have been invented as a specific family of logics for talking about conceptual knowledge .Although have started to investigate connections between these two families of logics there are still two different areas of research with lim- ited interaction. Both formalisms also play a central role in Data Semantics on the Web as the basis for languages like OWL and RDF. However, it has been recognized by different researchers that the special characteristics of Web Data often requires an interpretation that goes beyond the abilities of classical logic. In particular, the notion of consistency that is quite central to logic-based formalisms for capturing data semantics has to be reinvestigated as inconsistency is rather a rule than an exception on the Web.

4.2 Distributional & Lexical Semantics

In parallel to logic-based approaches to data semantics, a completely different approach to semantics has been devel- oped in the area of language processing. As an answer to the infeasibility of capturing the semantics of natural language, less formal ways of defining and exploring the meaning of terms have been investigated in the area of lexical seman- tics Here, the meaning of terms is described in terms of their relations to other terms. While this idea is similar to that of logic-based ontologies, the relations used are not formally defined, but rather describe the use of a term in tex- tual resources. Using these relations terms can be defined and disambiguated via the related terms An even more light-weight approach to describe the semantics of terms in natural language is via co-occurrence with other terms. While this approach also referred to as the vector space model of semantics as the meaning of a document or a term is represented using a term vector, is a purely statistical one, it can also be seen as a simplified version of lexical seman- tics that only uses a single relation. Due to its simplicity and scalability, lexical and especially distributional semantics has become quite popular in information search and retrieval.

V. SEMANTICS DATA USING

The specific characteristics of Web Data often requires the explicit use of semantic models the data processing. This dis- tinguishes semantic data models on the Web from traditional settings where semantic models like conceptual schemas were primarily used for the design and the documentation of Data and did not play a central role in the actual pro- cess of

using the data afterwards. Foregoing an explicit use of semantic models is possible if the data is centralized and there is an agreement on the intended meaning and use of the data. This is not the case on the Web as argued above. Consequently, semantic models play an important role in data processing on the web. In particular, there are two basic oper- ations that have been shown to benefit from an explicit use of semantic models, namely the search for and the integration of distributed data sources.

5.1 Semantic Search

Finding information has been a central problem on the Web from its creation on. Meanwhile commercial search engines in particular Google provide excellent support for finding web pages and textual documents based on key- word matching and advanced ranking methods.

5.2 Integrating Data

The nature of Web Data makes data integration a central problem. On the conventional web, the integration is a purely technical one: heterogeneous contents co-exists in different formats and can be accessed through the same infrastruc- ture. The invention of semi-structured data description lan- guages, i.e. XML addresses the data integration problem at the syntactical level: Data is represented using the same for- mat enabling users to process data using the same tools. Models of data semantics comes into play when a syntactic integration is not sufficient, but an integration on a structural and semantic level is needed.

VI. RESEARCH DIRECTIONS

In this article, I have tried to summarize the main issues con- nected with research relating to data semantics on the web. Starting with a discussion of the special properties of Web Data that makes it unique I have provided a brief survey of current ideas and principles of generating, representing and using data semantics on the Web. As the topic is a very broad and popular one, a lot of research is being done and will be done on this issue. In this section, I want to provide a per- sonal perspective on some lines of work that are very prom- ising topics that either require further investigation to solve existing problems or that have a high potential for creating progress.

VII. CONCLUSIONS

In summary, Data semantics on the web is both a challenging research topic that needs ideas from different fields of com- puter science. It thus provides an opportunity to create radi- cally new approaches on the boundaries of disciplines and test results from fields such as databases, information retrieval and artificial intelligence in a new challenging setting, lead- ing to new research questions in the different areas. On the other hand, being able to capture and represent the Semantics of Data on the Web has a huge potential for advanced appli-
cations in an area that rapidly gains importance in almost all areas of business and society including electronic commerce, political discourse and scientific exchange. This combination of a long-term research challenge and practical significants makes Data Semantics on the Web a topic that promises to remain long-term relevance and is clearly set apart from short term hypes that come and go in the process of scientific dis- covery.

REFERENCES

1. Aleksovski Z, Klein M, ten Kate W, van Harmelen F (2006) Match- ing unstructured vocabularies using a background ontology. In: Staab S, Svátek V (eds) Managing knowledge in a world of net- works, 15th international conference, EKAW 2006, Podebrady, Czech Republic. Lecture notes in computer science, vol 4248, pp 182–197
2. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29
3. Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P (2003) The description logic handbook: theory, implementation, applications. Cambridge University Press, Cambridge.

AUTHORS PROFILE

M MARKANDEYULU Pursuing M.Tech(CSE) from Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur, A.P., India . Email:markyellow9@gmail.com.

JAYARAJU CHIKKALA, Pursuing M.Tech(CSE) from Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur, A.P., India . Email: *glorious.ch@gmail.com.*