Searching Semantic Data Warehouses

Alfredo Cuzzocrea
ICAR-CNR and University of Calabria, Italy

INTRODUCTION

The problem of *searching Semantic Data Warehouses* lies in the between of two leading and well-understood research areas, i.e. *Search Computing* (also extended by means of *semantic technologies* – e.g., (Ceri & Brambilla, 2010)) and *Semantic Data Warehouses* (e.g., (Spaccapietra et al., 2009)). This problem has been of renewed attention at now, due to the important applications that found on a typical *Semantics Data Warehouse* (SDW) architecture. Among these, relevant ones convey in the large family represented by the *Semantic Web* applications such as *Ontology-based Web Information Systems* (e.g., (Wache et al., 2001)), *RDF-based Complex Systems* (e.g., (Nejdl et al., 2003)), *Analytical Tools over Large Resource-based Systems* (e.g., (Cuzzocrea et al., 2011)), *Web Warehouses* (e.g., (Bonifati & Cuzzocrea, 2007)), and so forth.

In more detail, Search Computing is a novel discipline whose main goal consists in enabling a novel paradigm for modeling, defining, and devising so-called *search services* that may be deployed over a large spectrum of architectures, ranging from centralized ones to more probing distributed ones (e.g., *Cloud Infrastructures* (Dikaiakos et al., 2009)). Search Computing is a key component of modern *Web search engines*, such as *Google* and *Yahoo!*, where the main issue is represented by the problem of locating, annotating, indexing and retrieving Web information from the *deep Web* via simple natural language queries. This problem has been investigated by means of a plethora of approaches, but modern proposals, mainly relying on semantic technologies, seem to playing a leading role in this scientific area. Briefly, these proposals pursue the idea of exploiting semantic approaches (e.g., Ontologies, RDF) to model Web resources, and then using this information at query execution time in order to magnify the effectiveness of the Web search engine.

Semantic Data Warehouses, instead, follow the main paradigm of equipping layers of architectures of classical Data Warehouses (e.g., (Chaudhuri & Dayal, 1997)), namely *ETL* (*Extraction*, *Transformation*, and *Loading*), *Storage*, and *DW Data Deployment/Presentation* (e.g., *OLAP* (Gray et al., 1997)) by means of semantic-inspired methodologies. Indeed, thanks for the evident synergies that this paradigm exposes with the related area of semantic-based Web resource modeling, annotating and discovering (like in early research experiences), the so-called *Semantic Web Data Warehousing* (e.g., (Binh et al., 2003)) seems to be one of the most promising initiative in the context investigated by our research. Indeed, Semantic Data Warehousing research opens of a widespread collection of research challenges that can reasonably be considered as "hot topics" for actual Database and Data Warehousing research. Among these, some that deserve recall are: the critical problem of defining *semantic-inspired methodologies for supporting ETL* (e.g., (Pequeno & Aparício, 2005)); *semantic-inspired techniques for annotating internal repositories of Data Warehouse storage layers* (e.g., (Feng & Dillon, 2003)); *semantic OLAP* (e.g., (Lakshmanan et al., 2002)), and so forth. From this, it clearly follows the enormous potentialities of Semantic Data Warehousing research in future years, even with important industrial applications (e.g., (Chute et al., 2010; McCusker et al., 2009)).

A first, important conclusion of our research is represented by the evident inadequateness of actual Data Warehouse architectures (e.g., (Chaudhuri & Dayal, 1997)) in supporting a whole semantic-based DW process, from the ETL phase to the knowledge-providing phase (e.g., OLAP, *reporting*, *Data Mining algorithms* etc). Indeed, by looking at traditional Data Warehouse architectures, contrary to deep investigation that, during these decades, has been done in the context of *Data Warehousing query optimization* (e.g., (Cuzzocrea et al., 2004; Cuzzocrea, 2005; Cuzzocrea et al., 2005; Cuzzocrea & Serafino, 2009)), the usage of semantics-based methodologies within their core layers is clearly limited and only related to model-

ing metadata of the storage layer (e.g., (Pequeno & Aparício, 2005)), to a narrowed extent. Some sporadic initiatives concern with the usage of semantics in the loading phase of specialized data sources, such as *XML data* (e.g., (Abiteboul et al., 2001)) and *sensor data* (e.g., (Ibrahim et al., 2005)). Contrary to this, the guidelines drawn by *Semantic Databases* (e.g., (Hull & King, 1987)) the major scientific area overlying Semantic Data Warehouse research, propose a more systematic usage of semantics-based methodologies in all the phases of the Data Warehousing process, from the integration of operational data sources (e.g., relational databases) to front-end tools (e.g., OLAP).

On the other hand, it is clear enough that dealing with searching Semantic Data Warehouses involves in the overlying problem of modeling, building and managing Semantic Data Warehouses, as a first step for designing and developing effective and efficient search algorithms over these novel data-intensive platforms.

Inspired by these motivations, this article provides the following contributions:

- An overview of related work falling in the context of Search Computing with semantics extensions;
- An overview of related work falling in the context of Semantic Data Warehouses and OLAP;
- Open issues and future directions in the searching Semantic Data Warehouses research area.

BACKGROUND

The term "Search Computing" embraces a wide collection of methodologies for enriching the capabilities of search engines in both effectiveness and expressive power (e.g., (Ceri & Brambilla, 2010)). In particular, among this wide family, those approaches that may have points in common with the typical phases of Data Warehouses (i.e., ETL, storage and knowledge-fruition) and that make use of a semantics-inspired paradigm are of particular interest.

Fazzinga and Lukasiewicz (2010) provides a quite-extended overview on approaches focusing on *semantic search over the Web*, a natural application context for

semantic search methodologies. Still in this line of research, (d'Amato et al., 2010) proposes models and algorithms for combining semantic Web search with *inductive reasoning* to improve the effectiveness and the expressive power of semantic Wed search engines and Fazzinga et al. (2011) extends the proposed framework to more relying *conjunctive queries*, while Mustapha et al. (2010a, 2010b) focuses on mixed *case-based reasoning* and Ontologies to achieve the same goal. *Documents*, along with Web pages, are also popular targets for semantic search techniques. In this context, Bikakis et al. (2010) proposes the integration of *keyword search* and semantics methods to achieve novel capabilities in searching large document corpuses, also in the more probing case of dealing with heterogeneous domains. A similar application scenario is investigated in Mrabet et al. (2010). Another interesting scientific area where semantic search plays a leading role is *Web Services* (e.g., (Paolucci et al., 2002)). Here, semantic techniques are used to *discover, matching and composing Web services in dynamic environments* like Clouds (e.g., (Armbrust et al., 2010)). Similar approaches have been proposed in the context of *complex systems*, such as *Wiki systems* (e.g., (Han & Zhao, 2010; Orlandi & Passant, 2010)). Still to highlight the widespread usage of semantics-inspired techniques, Kohn et al. (2010) proposes using them over *unstructured data* in order to produce *new knowledge* from already-available knowledge bases, Ahn et al. (2010) provides guidelines for improving *analytical processes* over large information bases by means of these techniques, Tamma (2010) focuses on retrieving business knowledge based on these techniques, (Bollegala et al., 2011; Yang & Kitsuregawa, 2011) study how to exploit semantic search in *text processing/mining*, and del Val Noguera et al. (2011) discusses the case of *self-adaptive systems*.

These techniques have been applied in different environments, particularly in *peer-to peer networks* (e.g., (Huang et al., 2011; Li & Xie, 2010; Xu & Zhang, 2010; Xue et al., 2010)), *information networks* (e.g., (Giannopoulos et al., 2010; Giunchiglia et al., 2010)), *Grid and Cloud environments* (e.g., (Guo et al., 2011)), *RDF graphs* (e.g., (Zhu et al., 2002)), and so forth.

For what instead regards to Semantic Data Warehouses, there not exists a wide literature, and lot of work still need to be done. Some proposals focus on the

Related Content

Risk Management via Digital Dashboards in Statistics Data Centers
Atif Amin, Raul Valverdeand Malleswara Talla (2020). *International Journal of Information Technologies and Systems Approach (pp. 27-45).*
www.irma-international.org/article/risk-management-via-digital-dashboards-in-statistics-data-centers/240763

Hardware Design for Decimal Multiplication
Mário P. Véstiasand Horácio C. Neto (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 5455-5464).*
www.irma-international.org/chapter/hardware-design-for-decimal-multiplication/112996

A Graph-Intersection-Based Algorithm to Determine Maximum Lifetime Communication Topologies for Cognitive Radio Ad Hoc Networks
Natarajan Meghanathan (2018). *Encyclopedia of Information Science and Technology, Fourth Edition (pp. 6536-6545).*
www.irma-international.org/chapter/a-graph-intersection-based-algorithm-to-determine-maximum-lifetime-communication-topologies-for-cognitive-radio-ad-hoc-networks/184349

ICT Investments and Recovery of Troubled Economies
Ioannis Papadopoulosand Apostolos Syropoulos (2018). *Encyclopedia of Information Science and Technology, Fourth Edition (pp. 2337-2344).*
www.irma-international.org/chapter/ict-investments-and-recovery-of-troubled-economies/183946

Impact of PDS Based kNN Classifiers on Kyoto Dataset
Kailasam Swathiand Bobba Basaveswara Rao (2019). *International Journal of Rough Sets and Data Analysis (pp. 61-72).*
www.irma-international.org/article/impact-of-pds-based-knn-classifiers-on-kyoto-dataset/233598