World Wide Web (2011) 14:461–463
DOI 10.1007/s11280-011-0139-z

Special issue on querying the data web

Novel techniques for querying structured data on the web

Paolo Ceravolo · Chengfei Liu · Mustafa Jarrar · Kai-Uwe Sattler

Published online: 17 August 2011

The rapid growth of structured data on the Web has created a high demand for making this content more reusable and consumable. Companies are competing not only on gathering structured content and making it public, but also on encouraging people to reuse and profit from this content. Many companies have made their content publicly accessible not only through APIs but also started to widely adopt web metadata standards such as XML, RDF, RDFa, and microformats.

This trend of structured data on the Web (Data Web) is shifting the focus of Web technologies towards new paradigms of structured-data retrieval. Traditional search engines cannot serve such data as the results of a keyword-based query will not be precise or clean, because the query itself is still ambiguous although the underlying data is structured. On the other side, traditional structured querying languages cannot be used directly as data on the Data Web is heterogeneous, large, distributed, schema-free, and not intuitive for web users. To expose the massive amount of structured data on the Web to its full potential, people should be able to query and combine this data easily and effectively.

P. Ceravolo (✉)
Via Mulini 5/B, Crema, CR 26013, Italy
e-mail: ceravolo@dti.unimi.it

P. Ceravolo
Università degli Studi di Milano, Milan, Italy

C. Liu
Swinburne University of Technology, Melbourne, Australia

M. Jarrar
Birzeit University, Bir Zeit, Palestine

K.-U. Sattler
Ilmenau University of Technology, Ilmenau, Germany

This special issue on "Querying the Data Web" was launched to explore new research trends and recent solutions to these challenges. There were 41 submissions and 8 papers were accepted for publication.

In the first paper "Distributed Processing of Continuous Sliding-Window k-NN Queries for Data Stream Filtering", Pripuzic, Zarko, and Aberer present their approach for processing sliding-window-based k-nearest neighbor (k-NN) queries. The main goal of this work is to support settings with multiple data streams as sources distributed over the Internet. For this purpose, the authors have designed a k-NN processing system on top of a CAN-based P2P system which allows to balance the load. In the article, they present the approach as well as results from an experimental evaluation showing the feasibility and scalability of the solution.

The second paper by Umbrich et al., "Comparing Data Summaries for Processing Live Queries over Linked Data", addresses the problem of querying Linked Data distributed over multiple sources. It focuses on using data summaries accessed during query evaluation for determining relevant sources to yield current query results. That is, each data source is summarized at the crawling time, and before answering a query, the query processor evaluates whether a source contains possible answers or should be excluded, i.e., source selection. This yields a faster query processing as only the relevant sources are considered. The article discusses several algorithms for generating these summaries, the lookup and joining strategies, and provides a comprehensive evaluation.

The third paper by Weixiong Rao and Lei Chen, "A Distributed Full-Text Top-K Document Dissemination in Distributed Hash Tables ", proposes a technique that gets good results in reducing the unit-publishing cost of syndication services, this is achieved through a DHT-based P2P overlay networks organized according to a keyword filtering structure.

The fourth paper by Lu et al., "Efficient Top-K Approximate Searches Against a Relation with Multiple Attributes", studies the problem of efficiently identifying top-K records that are most similar to a search query in the context of a relation with multiple attributes, and proposes two main approaches that identify the top-K records without computing similarity scores of all records.

The fifth paper by De Virgilio et al., "A Scalable and Extensible Framework for query answering over RDF", presents a novel approach for storing and querying RDF data in relational databases. The idea is based on the notion of construct, which represents a concept of the domain of interest. Constructs are represented differently at the conceptual, logical, and physical levels. An experiment using the Wikipedia dataset (about 47 Million Triples) shows that this approach yields a better performance than SDB Jena, Sesame, Openlink Virtuoso and OWLim systems.

The sixth paper by Bruno et al., "Indexing and Querying Segmented Web Pages: The BlockWeb Model", presents a model for indexing and querying Web pages. The main idea of this approach is to decompose pages into hierarchical block structures which are indexed and returned to queries instead of whole pages. The authors describe the transformation of Web into this model as well as techniques for indexing blocks in an XML repository. Experimental evaluations using two different datasets show that with this technique the mean average precision for retrieval can be improved.

The seventh paper by Fausto Giunchiglia et al., "Semantic Flooding Semantic Search across Distributed Lightweight Ontologies ", proposes a method to construct a semantic overlay network using classifications trees enriched with semantic links. The main

contribution of this work is related to the good accuracy that can be achieved with a small number of queries.

The eighth paper by Liu and Chen, "Processing keyword search on XML: a survey", reviews the works in the literature on XML keyword search from several perspectives, including defining relevant matches, defining relevant non-matches, ranking, result analysis, result evaluation, and indexes and materialized views. This survey performs a comprehensive analysis of the state-of-the-art techniques for processing keyword searches on XML.

We gratefully acknowledge the strong research community that gathered around the research problems related to querying the data web and the high quality of their research work, which is hopefully reflected in the papers of this special issue. We also would like to express our deep appreciation for the referees' hard work and dedication. Above all, thanks are due the authors for submitting the best results of their work to the WWWJ Special Issue on Querying the Data Web.