Special issue on structured and crowd-sourced data on the Web

Marco Brambilla · Stefano Ceri · Alon Halevy

Received: 25 June 2013 / Accepted: 26 June 2013 / Published online: 9 July 2013
© Springer-Verlag Berlin Heidelberg 2013

The abundance of structured and social data on the Web coupled with ability to solicit feedback from crowds has the potential of changing the way we search for information and enabling new classes of applications on the Web. This special issue of the VLDB Journal features original contributions that advance the state of the art in this topic area. Broadly, the special issue is concerned with methods for analyzing and serving structured data on the Web and methods for enhancing data by soliciting feedback from crowds.

Structured data appear on the Web in several forms, including hidden Web sources exposed through HTML form interfaces, tables, lists, and pages with repeating semistructured cards. Current research efforts for leveraging this data include approaches for extracting and combining results from multiple sources, for surfacing the deep web, and for exposing data through RDF repositories with rich linking, possibly exploiting existing knowledge bases for data annotation and integration.

Crowd-sourcing and social data are increasingly popular methods for improving search results and enhancing the quality of data on the Web. Social data have huge potential to re-rank and enrich pages and content based on what the user's friends have visited or recommended previously. Crowd-sourcing can be used to answer questions that are inherently hard for machines but can be handled relatively

easily with human input. The main challenges in these areas concern the quality assessment of the additional signals and blending socially promoted results with results generated by traditional algorithms.

A total of seventeen papers were submitted to the special issue; out of them, six papers were selected (acceptance rate 0.35); the paper by Bozzon et al. was managed by an anonymous VLDB Journal editor. The special issue deadline was September 15, 2012; five of the accepted papers went through both a major and a minor revision, and were resubmitted on April 1, 2013 and on May 15, 2013; one accepted paper had just a minor revision. Final acceptance occurred on June 1, 2013.

We next present a brief summary of accepted papers.

The paper “Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents,” by Goasdoué et al., proposes a novel hybrid data model capturing the structural aspects of XML data and the semantics of RDF, thus supporting pure XML or RDF datasets, as well as RDF-annotated XML data. As such, the approach enables managing a mix of semantic and purely syntactic content, which often occurs with Web data. In addition to the data model, the paper describes the XRQ query language, combining features of both XQuery and SPARQL, and experimentally assesses the proposed query processing algorithms in terms of time and quality.

The paper “The Ontological Key: Automatically Understanding and Integrating Forms to Access the Deep Web,” by Furche et al., presents a comprehensive approach to Web forms understanding and integration that covers both form labeling (combining features from the text, structure, and visual rendering of a web page) and form interpretation (based on a schema or ontology of forms in a given domain). The approach is validated through a lightweight form integration system that successfully translates and distributes

M. Brambilla (✉) · S. Ceri
Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, via Ponzio 34/5, 20133 Milan, Italy
e-mail: mbrambill@elet.polimi.it

S. Ceri
e-mail: ceri@elet.polimi.it

A. Halevy
Google Inc., 1600 Amphitheatre Parkway,
Mountain View, CA 94043, USA
e-mail: halevy@google.com

user queries to hundreds of forms. Experiments show that the approach achieves over 97 % accuracy in the evaluation domain.

The paper “Exploratory Search Framework for Web Data Sources,” by Bozzon et al., proposes a general-purpose exploratory search paradigm over Web data services. Exploratory search is an information-seeking behavior where users progressively learn about one or more topics of interest. In their work, the authors propose an exploratory user interface over schema and service-based data that include a set of widgets for data exploration, from big tables to atomic tables, visual diagrams, and geographic maps. User interactions are translated into queries defined in SeCoQL, a SQL-like language and protocol specifically designed for supporting exploratory search over data sources. Effectiveness of the approach is evaluated from the end-user perspective in the context of a cognitive model for search.

The paper “Large-Scale Linked Data Integration Using Probabilistic Reasoning and Crowd-sourcing,” by Demartini et al., considers the problem of semiautomatically matching large collections of Web pages to linked data. The authors propose a three-stage blocking technique in order to obtain the best possible instance matches while minimizing both computational complexity and latency. They identify entities from natural language text using the state of the art techniques and then automatically connect them to the Linked Open Data cloud. To improve performance, the matching algorithm first finds potential candidate results through structured inverted indices over already extracted entities and only then they refine the matches by querying a graph database, which is a more expensive operation. In case the automatic algorithms fail to come up with convincing results, they resort to human computation.

The paper “Schema Matching Prediction with Applications to Data Source Discovery and Dynamic Ensembling,” by Sagi and Gal, introduces schema matching prediction for Web-scale data integration. The technique provides assistance to human schema matchers in the absence of exact matches between schema elements. The approach is based on a predictor using similarity spaces that predicts the success of a matcher in identifying correct correspondences. The paper proposes a method for constructing and tuning predictors, and also studies the desirable properties of predictors, namely correlation, robustness, the ability to tune, and generalization.

The paper “Hybrid Entity Clustering Using Crowds and Data,” by Lee et al., addresses the problem of clustering query results at the entity level. Clustering at the entity level is more challenging than traditional document clustering because diverse similarity notions between entities need to be supported in heterogeneous domains. The paper proposes a hybrid machine- and crowd-based relationship-clustering algorithm that exploits co-occurrence and numeric features. In particular, the proposed technique captures diverse user perceptions from co-occurrence and disambiguates different senses using feature-based similarity.

While the special issue offers some of the latest developments on Web data management and on crowd-sourcing, we see that the two fields are starting to come together in interesting ways. Crowd-sourcing is starting to be used to solve hard data integration problems that are common on the Web. Given the diversity, breadth, and subjectivity of Web data, we expect crowd-sourcing to play an even greater future role in leveraging this incredible resource.

June 20, 2013.